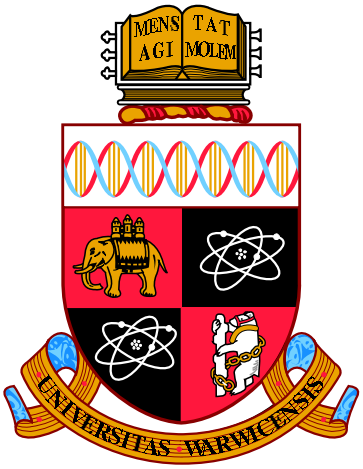


Optimal control and inverse problems involving point and line functionals and inequality constraints

by

Charles Brett

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Mathematics

July 2014

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank my supervisors Professor Charles Elliott and Dr Andreas Dedner for their support throughout my PhD. They suggested my original research project and their ideas have led me in many interesting research directions.

I am grateful to Professor Michael Hintermüller and Dr Caroline Löbhard for a fruitful collaboration, leading to a chapter in this thesis. It gave me the opportunity to visit Berlin, which I enjoyed immensely. I am also grateful for the feedback provided by Professor Andrew Stuart during my annual Personal Advisory Committee meetings.

Finally, thank you to various members of the mathematics department at the University of Warwick, and in particular my fellow MASDOC PhD students, for creating an exciting research environment.

This research was made possible by the financial support of the UK Engineering and Physical Sciences Research Council (EPSRC) through Grant EP/H023364/1.

Declarations

Chapter 1 introduces standard notation and well known results from the literature on the optimal control of PDEs.

Chapters 2 and 3 contain original work, which I proved with the assistance of my supervisors Professor Charles Elliott and Dr Andreas Dedner.

Chapter 4 contains original work that resulted from a collaboration with Professor Michael Hintermüller and Dr Caroline Löbhard from Humboldt Universität Berlin. This work is contained in a paper that has been submitted for publication in [Brett et al., 2013]. Although there was input in both directions for all parts of this work, the analytical results in Sections 4.2 and 4.3 were mainly proved by Dr Caroline Löbhard, and so are stated without proof. The discretisation and derivation of the estimator in Sections 4.4 and 4.5 was done jointly. Section 4.6 until the end of the chapter is mainly my own work.

Chapter 5 contains original work that I completed with the support of my supervisors Professor Charles Elliott and Dr Andreas Dedner. It has been accepted for publication in [Brett et al., 2014]. Note that the abstract framework and theoretical results extend previous work of my supervisor Professor Charles Elliott. The comparison of the double well and double obstacle potential has not been done previously for the type of problem under consideration.

The work on the materials science application in Section 5.8 is original and is joint work with Dr Nils Warnken from the University of Birmingham.

Abstract

In this thesis we consider some problems related to the optimal control of partial differential equations (PDEs) and variational inequalities (VIs) with various constraints. Such problems are important because in real world applications we are typically more interested in optimising and controlling processes than just simulating them. We focus on developing efficient solution methods for these problems.

The first part of this thesis considers optimal control of PDEs and VIs but with the usual L^2 fidelity term replaced by ones which encourages the state to take certain values at points or along surfaces of codimension 1. Such problems are related to optimal control with pointwise state constraints, which are relevant in applications. Our new fidelity terms cause complications in the formulation of the optimal control problems, as well as the analysis and the numerical analysis.

The second part of this thesis considers the inverse problem of recovering a binary function from blurred and noisy data. Such image processing problems arise in many applications, for example decoding barcodes. Our approach uses the Mumford-Shah model, but with a phase field approximation to perimeter regularisation. We develop iterative methods for solving the problem and prove convergence results. Numerical results are presented which illustrate the effectiveness of our approach and the relative merits of different phase field approximations. We finish by applying our algorithms to a problem in materials science.

Chapter 1

Introduction

Many processes in fields such as physics, engineering and finance can be modelled by partial differential equations (PDEs) and variational inequalities (VIs), for example heat conduction, fluid flows, and crystal growth. In real world applications the end goal is generally not the modelling and simulation of processes; we want to use models to design, optimise and control them. In particular, we often want to consider optimal control and inverse problems. The focus of this thesis is on the numerical solution of such problems. This is challenging as after discretisation they typically have a very large number of optimisation variables. Only recently have advances in numerical methods and computing power made it practical to solve them, and even then a thorough mathematical analysis is required in order to design efficient solution methods.

In this introductory chapter we will explain the basic terminology used for optimal control of PDEs, and touch upon their link to inverse problems. This may not be familiar to readers with a background in just PDEs or numerical analysis. It will also allow us to highlight where the problems we consider in this thesis deviate from standard ones. For a more detailed introduction to the optimal control of PDEs see e.g. [Tröltzsch, 2010; Hinze et al., 2009].

Consider the simple elliptic PDE

$$\begin{aligned} -\Delta y &= \eta \quad \text{in } \Omega, \\ y &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{1.1}$$

Such a PDE may arise, for example, when modelling the long term temperature distribution of an object with a fixed boundary temperature. In this situation η could represent a distributed heat source and y the temperature distribution. We may want to consider the problem of finding the heat source that gives a particular

temperature distribution g_d. Then we would call the source term η the *control*, the solution y the *state*, g_d the *desired state*, and (1.1) the *state equation*.

For the PDE (1.1), if $u \in L^2(\Omega)$ then in general there is no solution in $C^2(\Omega)$ and the PDE cannot be solved in the classical sense. We may want to allow controls with this regularity, so it is typical to instead work with a weak formulation of the state equation i.e. find y in the Sobolev space $H_0^1(\Omega)$ such that

$$(\nabla y, \nabla v)_{L^2(\Omega)} = (\eta, v)_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega). \tag{1.2}$$

We call the operator that maps the control η to the state y satisfying (1.2) the *control-to-state* operator, and denote it by S. In this example we choose the domain and range of S (the *control space* and *state space*) to both be $L^2(\Omega)$, but other choices will be used in the later chapters.

If the desired state g_d is in $L^2(\Omega)$ there may be no control η such that $S\eta = g_d$, for example if S has a smoothing effect. We can instead try and find a control such that the state is close to g_d in some sense. Also, in many applications the control may have some cost associated to it, so we want a compromise between the state being close to g_d and the control being small in some sense. We can achieve both of these aims by minimising an *objective functional* that depends on the control and the state, subject to (s.t.) the constraint that the state equation must hold. An objective functional that is often used is $J : L^2(\Omega) \times L^2(\Omega) \to \mathbb{R}$ defined by

$$J(y, \eta) = \frac{1}{2}\|y - g_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)}.$$

The first term in J is sometimes referred to as the *fidelity term*, as it penalises deviation of the state from the desired state. The second term is sometimes referred to as the *cost term*, and it is weighted by the *cost of the control* ν, a nonnegative real number that allows for different relative weights between the fidelity and cost terms.

We may also want to impose further constraints on the problem. For example, we cannot have an infinite heat source so we could include the *control constraint* that the control belongs to U_{ad}, a subset of the control space called the *set of admissible controls*. A typical form for this is

$$U_{ad} := \{\eta \in L^2(\Omega) : a \leq \eta(x) \leq b \text{ a.e. } x \text{ in } \Omega\}, \tag{1.3}$$

where $a, b \in \mathbb{R}$ with $a < b$. Similarly we could impose *state constraints*, but the problems we consider in this thesis do not have these.

To summarise, the optimal control problem we have just introduced is

$$\begin{aligned} &\min\ J(y,\eta) \\ &\text{over } L^2(\Omega)\times L^2(\Omega) \\ &\text{s.t. } y = S\eta \\ &\text{and } \eta \in U_{ad}. \end{aligned} \tag{1.4}$$

Note that although optimal control problems are often written like this, what we actually want to find is a control $\eta \in L^2(\Omega)$ with state $y = S\eta$ such that (y,η) minimises J; we are not just trying to find the smallest value of J. We call such a control an *optimal control* and typically denote it by u.

We can use the control-to-state operator to define the *reduced objective functional* $\hat{J} : L^2(\Omega) \to \mathbb{R}$ by

$$\hat{J}(\eta) = J(S\eta, \eta), \tag{1.5}$$

which just depends on the control. It is straightforward to see that (Su, u) minimises (1.4) if and only if u is the minimiser of the following optimisation problem:

$$\min\ \hat{J}(\eta) \text{ over } \eta \in U_{ad}. \tag{1.6}$$

We will also refer to problems in this equivalent form as optimal control problems.

Before commenting on existence of a solution to this problem we introduce a more abstract formulation, which the above example fits into, as this highlights the mathematical assumptions we are making. Let the control space U be a Hilbert space and let U_{ad} be a nonempty, closed and convex subset of U (but not necessarily as defined in (1.3)). Let the state space H be a Hilbert space and take $S : U \to H$ to be a continuous linear operator i.e. $S \in \mathcal{L}(U,H)$. Then for $\nu \geq 0$ redefine $\hat{J}$ as

$$\hat{J}(\eta) = \frac{1}{2}\|S\eta - g_d\|_H^2 + \frac{\nu}{2}\|\eta\|_U^2$$

and consider the optimal control problem (1.6). It is well known that this problem has a solution (see e.g. Theorem 2.14 in [Tröltzsch, 2010]). Furthermore if either $\nu > 0$ or S is injective, then $\hat{J}$ is strictly convex and the optimal control is unique.

Note that $\hat{J}$ is a convex Gâteaux differentiable functional with Gâteaux derivative $\hat{J}' : U \to U^*$ (where U^* denotes the dual space of U). Moreover U_{ad} is convex. Therefore another well known result (see e.g. Lemma 2.21 in [Tröltzsch, 2010]) says that u is an optimal control for (1.6) if and only if the following varia-

tional inequality holds:

$$u \in U_{ad}, \quad \langle \hat{J}'(u), v - u \rangle_{U^*} \geq 0 \quad \forall v \in U_{ad}, \tag{1.7}$$

where $\langle \cdot, \cdot \rangle_{U^*}$ denotes the usual duality pairing between U^* and U. We call a necessary and sufficient condition such as (1.7) an *optimality condition*. Sometimes $\hat{J}$ may not be convex, in which case (1.7) is only a necessary condition and a u satisfying it may not be an optimal control. In this case we refer to (1.7) as a *stationarity condition* and a function satisfying it as a *stationary point*. Note that if there is no control constraint (i.e. $U_{ad} = U$) then (1.7) simplifies to the equality

$$\hat{J}'(u) = 0 \text{ in } U^*.$$

We can use the Riesz representation theorem to write (1.7) as

$$u \in U_{ad}, \quad (Su - g_d, S(v - u))_H + \nu(u, v - u)_U \geq 0 \quad \forall v \in U_{ad}, \tag{1.8}$$

where $(\cdot, \cdot)_H$ denotes the inner product on H and similarly for $(\cdot, \cdot)_U$. Making use of the *adjoint operator* $S^* : H \to U$ of S, which is defined by

$$(S^* p, v)_U = (p, Sv)_H \quad \forall p \in H, v \in U,$$

observe that we can introduce the variable $p = S^*(Su - g_d) \in U$ in the above inequality. This allows (1.8) to be equivalently written as:

$$\begin{aligned} &y = Su, \\ &p = S^*(y - g_d), \\ &u \in U_{ad}, \quad (p + \nu u, v - u) \geq 0 \qquad\qquad \forall v \in U_{ad}. \end{aligned}$$

We call p the *adjoint variable*. It can also be thought of as a *Lagrange multiplier* for the constraint that the state equation holds. Even though we have increased the number of unknowns, it is often more efficient to solve this system for (y, u, p) simultaneously, rather than directly solve (1.8) for u.

Returning to our particular example of (1.4) with $H = U = L^2(\Omega)$, S defined by (1.2) and U_{ad} given by (1.3), the optimality conditions become a system of coupled

PDEs and a variational inequality:

$$\begin{aligned}
&(\nabla y, \nabla v)_{L^2(\Omega)} = (u, v)_{L^2(\Omega)} && \forall v \in H_0^1(\Omega),\\
&(\nabla p, \nabla v)_{L^2(\Omega)} = (y - g_d, v)_{L^2(\Omega)} && \forall v \in H_0^1(\Omega),\\
&u \in U_{ad}, \quad (p + \nu u, v - u) \geq 0 && \forall v \in U_{ad}.
\end{aligned}$$

In order to solve optimal control problems numerically there are two possible approaches. We could discretise the optimal control problem then optimise (i.e. replace the objects in the optimal control problem by discrete objects then derive optimality or stationarity conditions for this discrete optimal control problem). Alternatively we could optimise then discretise i.e. derive optimality conditions in function space (as we did in (1.7)) then replace these conditions with a discrete version. For our problems we take the discrete space for the adjoint to be the same as the discrete space for the state, so these approaches coincide. Our discretisation is done using a finite element method, and in particular piecewise linear globally continuous finite elements.

We will now briefly mention how the problems in the rest of this thesis differ from the above standard example of an optimal control problem. This will be elaborated on in the introduction to each chapter.

In Chapter 2 we consider the above optimal control problem but with a new fidelity term which penalises deviation of the state from prescribed values at a finite set of prescribed points. This problem is related to optimal control problems with state constraints. In order for it to be well posed we require the control-to-state operator to map to continuous functions, which affects the analysis and numerical analysis of the problem. In Chapter 3 we consider a fidelity term which penalises deviation of the state from a function along a surface of codimension 1 (i.e. a line in two dimensions and a surface in three dimensions). This does not change the analysis significantly but it affects the discrete problem and therefore its numerical analysis. In Chapter 4 we consider the point fidelity term again but this time the state equation is replaced by a variational inequality. This control-to-state operator is nonlinear, resulting in a difficult nonconvex and nondifferentiable optimisation problem.

Optimal control problems have close links to inverse problems like the one we consider in Chapter 5. For example, we may want to determine some quantity u that can only be observed through noisy observations of a quantity Su. So we have observations $g_d = Su + \zeta$ for some unknown noise ζ and want to find u. Typically such problems are ill posed. Instead we could formulate the problem

(roughly speaking) as finding η minimising $\|S\eta - g_d\|_Y$ for some norm $\|\cdot\|_Y$. This problem may have many solutions so it is common to regularise it, for example by minimising

$$\frac{1}{2}\|S\eta - g_d\|_Y^2 + \frac{\nu}{2}R(\eta)$$

for $\nu > 0$, where perhaps $R(\eta) = \|\eta\|_Z^2$ for some norm $\|\cdot\|_Z$. Therefore this inverse problem has the same mathematical structure as the optimal control problem we introduced earlier. In Chapter 5 we formulate the inverse problem of binary image recovery precisely and then investigate numerical methods for solving it. In this case S will be a blurring operator (such as the solution operator of an elliptic PDE) and the regularisation will be the total variation. We apply our numerical methods for solving this problem to an application in materials science in Section 5.8.

Each chapter in this thesis will follow a similar structure. We first describe the problem under consideration and motivate its study. We then analyse the problem, in particular commenting on existence and uniqueness of solutions and deriving optimality or stationarity conditions. Next we discretise the problem using finite elements and attempt some numerical analysis. Depending on the problem we may be able to prove error estimates, just convergence, or perhaps only numerically justify the effectiveness of a solution algorithm.

Chapter 2

Optimal control of elliptic PDEs at points

In this chapter we study an elliptic optimal control problem with an objective functional containing the distance between the state and prescribed values at a finite number of prescribed points. This contrasts with standard elliptic optimal control problems, where typically the objective functional contains the L^2 distance between the state and the desired state over the whole domain. So for a bounded domain $\Omega \subset \mathbb{R}^n$ ($n = 2$ or 3) with boundary $\partial\Omega$ we consider the problem:

$$\min \frac{1}{2} \sum_{\omega \in I} (y(\omega) - g_\omega)^2 + \frac{\nu}{2} \|\eta\|^2_{L^2(\Omega)}$$

subject to the state equation

$$\begin{aligned} Ay &= \eta \quad \text{in } \Omega \\ y &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{2.1}$$

and the control constraints

$$a \leq \eta \leq b.$$

Here $I \subset \Omega$ is a finite set of points, $\{g_\omega\}_{\omega \in I} \subset \mathbb{R}$ are prescribed values at these points, $\nu > 0$ is the cost of control, A is an elliptic operator, and $a, b \in \mathbb{R}$ with $a < b$ are lower and upper bounds for the control. We give the precise statement of the problem using function spaces in Section 2.2.

The motivation for the point fidelity term is that in some applications we may only care about the state being close to given values at certain points in the domain. Controlling the state using a distributed norm over the whole domain

yields weaker control at points. The point fidelity term encourages the state to take certain values at points, so our problem is closely related to one which imposes hard constraints on the state at points. Imposing hard state constraints can often lead to an optimal control with a very high cost, whereas our point fidelity term allows for a compromise between how close the state is to the prescribed values and the cost of the control. On the other hand, we will prove later that as we increase the relative weighting given to the point fidelity term, the solutions of point control problems converge weakly to the solution of a problem with point state constraints.

In the literature there are computational results for PDE optimal control problems with objective functionals that contain point evaluations of the state. However we have not found any literature that contains a numerical analysis of such problems. [Tröltzsch, 2010] formulates an optimal control problem where the objective functional is the state evaluated at a point, but does not discuss numerical methods for solving it. [Unger and Tröltzsch, 2001] considers optimally controlling the cooling of steel. This problem is formulated with an objective functional that contains the temperature of the steel at a number of points (i.e. point evaluations of the state) as this makes the problem more tractable. The paper focuses on computational results and the numerical analysis is not considered. The medical imaging problem of electrical impedence tomography (see e.g. [Hintermüller and Laurain, 2008]) could be formulated as an inverse problem with a point fidelity term (but with the points on the boundary). This is because one reconstructs a conductivity based on measurements of the voltage over small regions, which could be approximated by measurements at points. In the paper [Brett et al., 2013] (written by ourselves) the point fidelity term is used for the optimal control of elliptic variational inequalities. The difficulty of the nonlinear control-to-state operator means that an a posteriori error estimator is derived but a priori error estimates are not considered.

Our aim is to fill a gap in the literature by studying in detail the numerical analysis of a finite element approximation of the above point control problem, which could be considered the canonical optimal control problem with an objective functional containing point evaluations of the state. However related problems have been considered in the literature. The recent paper [Gong et al., 2014] considers elliptic optimal control problems with controls at points and on other lower dimensional manifolds. The numerical analysis of these problems leads to mathematical difficulties similar to those in this chapter. In particular, when the control is at points the state equation has delta functions on the right hand side, where as in our problem the adjoint equation has delta functions. In both cases this means low

regularity of the state/adjoint. In the thesis [Brett, 2014] theory is developed for an elliptic optimal control problem where the fidelity term is an integral along a surface of codimension 1, which is also a set of measure zero relative to the domain. In papers such as [Casas et al., 2012] and [Pieper and Vexler, 2013] elliptic optimal control problems are considered where the control spaces are spaces of measures.

Regularity issues are also faced by elliptic optimal control problems with state constraints. [Leykekhman et al., 2013] proves error estimates for problems with state constraints at a finite number of points. Note that this paper also proves improved error estimates for graded triangulations (such triangulations are locally refined towards the singularities but have asymptotically the same number of elements for a given triangulation size), but we do not consider these. [Deckelnick and Hinze, 2007] proves error estimates for the case of global (as opposed to point) state constraints, but for a state equation with Neumann boundary conditions. Parabolic optimal control problems often contain point evaluations in time of the state, but these are functions over the space domain and the technicalities of the numerical analysis are different. A review of the analysis for standard elliptic and parabolic optimal control problems can be found in [Tröltzsch, 2010] and a review of the numerical analysis can be found in [Hinze et al., 2009].

In this chapter we use two different methods of discretising our problem with finite elements. The first method is to explicitly discretise the control by minimising over a space of discrete controls, leading to discrete problem ($\mathrm{M1}_h$) (see (2.31)). The second method is to implicitly discretise the control through a discrete control-to-state operator using the variational discretisation concept of [Hinze, 2005], leading to discrete problem ($\mathrm{M2}_h$) (see (2.34)). We later observe that when there are no control constraints these two methods may lead to equivalent discrete problems. We are not able to prove an estimate for ($\mathrm{M1}_h$) in dimension 3 with control constraints, which motivates us to use ($\mathrm{M2}_h$) for our implementation despite it being less standard to solve computationally.

Next we use two different approaches to prove a priori error estimates for the $L^2(\Omega)$ error in the control for these discrete problems. The first approach (Approach 1, Section 2.4.1) is inspired by [Casas and Tröltzsch, 2003] and the second approach (Approach 2, Section 2.4.2) is inspired by [Deckelnick and Hinze, 2007]. The main estimates we prove are summarised in Table 2.1, where $\varepsilon > 0$ is arbitrary. We see that Approach 2 does not offer any better error estimates than Approach 1. However we include Approach 2 because it is simpler when it applies. Numerical results confirm that the error estimates are realised for ($\mathrm{M2}_h$).

In the next section we introduce some notation. In Section 2.2 we formu-

Discretisation	$(\mathrm{M1}_h)$	$(\mathrm{M1}_h) = (\mathrm{M2}_h)$	$(\mathrm{M2}_h)$
Dimensions	$n = 2$	$n = 2, 3$	$n = 2, 3$
Constraints	both	$b = -a = \infty$	both
Approach 1	$O(h)$	$O(h^{2-\frac{n}{2}})$	$O(h^{2-\frac{n}{2}})$
Approach 2	-	$O(h^{2-\frac{n}{2}-\varepsilon})$	$O(h^{2-\frac{n}{2}-\varepsilon})$
Numerics	-	$O(h^{2-\frac{n}{2}})$	$O(h^{2-\frac{n}{2}})$

Table 2.1: The main a priori error estimates proved for $\|u - u_h\|_{L^2(\Omega)}$.

late the optimal control problem precisely and prove some analytical results. In Section 2.3 we discretise using the finite element method. In Section 2.4 we prove a priori error estimates for the L^2 error in the control. In Section 2.5 we show numerical results.

2.1 Notation

We begin by introducing some function spaces that are needed to formulate the optimal control problem precisely.

Let the domain $\Omega \subset \mathbb{R}^n$ ($n = 2$ or 3) be a bounded open set that either has a smooth boundary or is convex with a polygonal (for $n = 2$) or polyhedral (for $n = 3$) boundary. Both $C(\bar{\Omega})$ and its subspace $C_0(\Omega)$ (of functions that are zero on $\partial\Omega$) are Banach spaces when endowed with the supremum norm, $\|\cdot\|_\infty$. Recall that for $n = 2$ or 3, the Sobolev space $H^2(\Omega)$ is continuously embedded into $C(\bar{\Omega})$ (see e.g. [Adams and Fournier, 2003]), so $H^2(\Omega) \cap H_0^1(\Omega) \subset C_0(\Omega)$. By different versions of the Riesz Representation Theorem (see e.g. Theorems 2.14 and 6.19 in [Rudin, 1987]) the dual spaces of $C(\bar{\Omega})$ and $C_0(\Omega)$ can both be identified with the space $\mathcal{M}(\Omega)$ of real regular Borel measures on Ω. In particular, for $\mu \in \mathcal{M}(\Omega)$ and $v \in C(\bar{\Omega})$ define the duality pairing

$$\langle \mu, v \rangle_{\mathcal{M}(\Omega)} := \int_\Omega v \mathrm{d}\mu,$$

where the integral is the Lebesgue integral with respect to μ. Here $\langle \mu, v \rangle_{\mathcal{M}(\Omega)}$ abbreviates $\langle \mu, v \rangle_{\mathcal{M}(\Omega), C(\bar{\Omega})}$. Then for each $z \in C(\bar{\Omega})^*$ there exists a unique $\mu \in \mathcal{M}(\Omega)$ such that

$$z(v) = \langle \mu, v \rangle_{\mathcal{M}(\Omega)} \quad \forall v \in C(\bar{\Omega}). \tag{2.2}$$

The same result holds for $z \in C_0(\Omega)^*$ using the same definition of $\langle \mu, v \rangle_{\mathcal{M}(\Omega)}$ but with $v \in C_0(\Omega)$. We prefer to write $\int_\Omega v \mathrm{d}\mu$ but will sometimes use $\langle \mu, v \rangle_{\mathcal{M}(\Omega)}$ to

simplify notation. Note that $\mathcal{M}(\Omega)$ is a Banach space with the norm

$$\|\mu\|_{\mathcal{M}(\Omega)} := |\mu|(\Omega) = \sup\Big\{\int_\Omega v \mathrm{d}\mu : v \in C_0(\Omega) \text{ and } \|v\|_\infty \leq 1\Big\},$$

where $|\mu|$ is called the total variation of μ. For example, the Dirac measure centred at a point $\omega \in \Omega$, which we denote by δ_ω, is contained in $\mathcal{M}(\Omega)$ and $\|\delta_\omega\|_{\mathcal{M}(\Omega)} = 1$.

We will need the following embedding results for the Sobolev spaces $W_0^{1,s}(\Omega)$, where $V \hookrightarrow W$ denotes that V is continuously embedded into W.

Remark 2.1. *From [Adams and Fournier, 2003] we have that:*

- *For $s > n$, $W^{1,s}(\Omega) \hookrightarrow C(\bar{\Omega})$;*
- *For $s > \frac{2n}{n+2}$, $W^{1,s}(\Omega) \hookrightarrow L^2(\Omega)$;*
- *For $s < \frac{2n}{n-2}$, $H^2(\Omega) \hookrightarrow W^{1,s}(\Omega)$.*

Consider the Dirichlet problem (2.1), where the differential operator A acting on a function $z : \Omega \to \mathbb{R}$ is defined by

$$Az = -\sum_{i,j=1}^{n} \partial_{x_j}(a_{ij}\partial_{x_i} z) + a_0 z$$

with

$$\begin{aligned}
&a_0 \in L^\infty(\Omega), \quad a_0(x) \geq 0 \quad \text{ for a.e. } x \in \Omega,\\
&a_{ij} = a_{ji} \in C^1(\bar{\Omega}),\\
&\exists\, \alpha > 0 \text{ s.t. } \sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \alpha |\xi|^2, \quad \forall x \in \Omega,\, \xi \in \mathbb{R}^n.
\end{aligned}$$

In particular, $A = -\Delta$ satisfies these assumptions. We want to work with a weak formulation of (2.1). Define the conjugate q' of q to be the real number such that $\frac{1}{q} + \frac{1}{q'} = 1$, and define the bilinear form $a : W_0^{1,q}(\Omega) \times W_0^{1,q'}(\Omega) \to \mathbb{R}$ associated to A by

$$a(z, v) = \sum_{i,j=1}^{n} \int_\Omega a_{ij}(x)\partial_{x_i} z(x)\partial_{x_j} v(x)\mathrm{d}x + \int_\Omega a_0(x)z(x)v(x)\mathrm{d}x,$$

where the derivatives are taken in the weak sense. By a standard result, for $\eta \in L^2(\Omega)$ there is a unique $y \in H_0^1(\Omega)$ satisfying

$$a(y, v) = (\eta, v) \quad \forall v \in H_0^1(\Omega). \tag{2.3}$$

Here and throughout this chapter $(\cdot,\cdot)$ denotes the $L^2(\Omega)$ inner product. With the regularity we assume on the boundary of Ω we have that $y \in H^2(\Omega) \cap H_0^1(\Omega)$ and

$$\|y\|_{H^2(\Omega)} \leq C\|\eta\|_{L^2(\Omega)}.$$

Here and throughout this chapter C is a positive constant that may vary from line to line and is independent of the variables it precedes (e.g. in the above equation C is independent of η). For a proof of this regularity and stability result see Theorems 2.2.2.3 and 3.2.1.2 in [Grisvard, 1985]. Since $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$ we in fact have that $y \in C_0(\Omega)$ and

$$\|y\|_\infty \leq C\|\eta\|_{L^2(\Omega)}. \tag{2.4}$$

We define the control-to-state operator $S : L^2(\Omega) \to C_0(\Omega)$ to map $\eta \in L^2(\Omega)$ to the solution $y \in C_0(\Omega)$ of (2.3). S is linear, and also continuous by (2.4), so S has an adjoint operator. Using (2.2) we can define the adjoint $S^* : \mathcal{M}(\Omega) \to L^2(\Omega)$ of S by

$$(S^*\mu, \eta) = \langle \mu, S\eta \rangle_{\mathcal{M}(\Omega)} \quad \forall \mu \in \mathcal{M}(\Omega), \eta \in L^2(\Omega). \tag{2.5}$$

Note that the control-to-state operator S has the following characterisation.

Lemma 2.2. *For $\eta \in L^2(\Omega)$, $y = S\eta$ if and only if $y \in C_0(\Omega)$ satisfies*

$$\forall q \in \left(n, \frac{2n}{n-2}\right) : \quad y \in W_0^{1,q}(\Omega), \quad a(y,v) = (\eta, v) \quad \forall v \in W_0^{1,q'}(\Omega). \tag{2.6}$$

Here (η, v) makes sense since $q \in (n, \frac{2n}{n-2})$ if and only if $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$, and Remark 2.1 gives that for such q' we have $v \in W_0^{1,q'}(\Omega) \subset L^2(\Omega)$.

Proof. Suppose $y = S\eta$ (i.e. it solves (2.3)) and take $q \in (n, \frac{2n}{n-2})$. Since $y \in H^2(\Omega)$ we can integrate $a(y,v)$ by parts against $v \in C_c^\infty(\Omega)$ to get

$$a(y,v) = (Ay, v) \quad \forall v \in C_c^\infty(\Omega). \tag{2.7}$$

Then using (2.3) we get

$$(\eta, v) = (Ay, v) \quad \forall v \in C_c^\infty(\Omega), \tag{2.8}$$

which implies that $Ay = \eta$ a.e. in Ω. Moreover, it follows from (2.7) and the density of $C_c^\infty(\Omega)$ in $W_0^{1,q'}(\Omega)$ that $a(y,v) = (Ay, v)$ for all $v \in W_0^{1,q'}(\Omega)$. Combining this fact, $Ay = \eta$ a.e. in Ω and $v \in W_0^{1,q'}(\Omega) \subset L^2(\Omega)$ gives $a(y,v) = (\eta, v)$ for all $v \in W_0^{1,q'}(\Omega)$. By Remark 2.1 note that $y \in H^2(\Omega) \cap H_0^1(\Omega) \subset W_0^{1,q}(\Omega)$. The above arguments hold for any $q \in (n, \frac{2n}{n-2})$, so we have proved that $y = S\eta$ implies (2.6)

holds.

The reverse implication is also true. Since $H_0^1(\Omega) \subset W_0^{1,q'}(\Omega)$ for any $q \in (n, \frac{2n}{n-2})$, we can test (2.6) with any $v \in H_0^1(\Omega)$. So a solution of this must solve (2.3). This completes the proof. □

We can use this result to prove that the adjoint operator S^* can be characterised in the following way.

Lemma 2.3. *For $\mu \in \mathcal{M}(\Omega)$, $p = S^*\mu$ if and only if $p \in L^2(\Omega)$ satisfies*

$$\forall q' \in \Big(\frac{2n}{n+2}, \frac{n}{n-1}\Big) : \quad p \in W_0^{1,q'}(\Omega), \quad a(v,p) = \int_\Omega v \,\mathrm{d}\mu \quad \forall v \in W_0^{1,q}(\Omega). \tag{2.9}$$

Moreover,

$$\|p\|_{W_0^{1,q'}(\Omega)} \leq C(q')\|\mu\|_{\mathcal{M}(\Omega)} \quad \forall q' \in \Big(\frac{2n}{n+2}, \frac{n}{n-1}\Big). \tag{2.10}$$

Proof. Suppose (2.9) is true. Fix some $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$ then for all $\mu \in \mathcal{M}(\Omega)$ and $\eta \in L^2(\Omega)$, testing (2.9) with $S\eta \in W_0^{1,q}(\Omega)$ gives

$$a(S\eta, p) = \int_\Omega S\eta \,\mathrm{d}\mu = \langle \mu, S\eta \rangle_{\mathcal{M}(\Omega)}.$$

By the definition of S we can test (2.6) with $p \in W_0^{1,q'}(\Omega)$ to get

$$a(S\eta, p) = (\eta, p) = (p, \eta).$$

Combining these two equalities and recalling that μ and η are arbitrary we get

$$\langle \mu, S\eta \rangle_{\mathcal{M}(\Omega)} = (\eta, p) \quad \forall \mu \in \mathcal{M}(\Omega), \eta \in L^2(\Omega).$$

Comparing this to the definition of the adjoint we see $p = S^*\mu$. Since q' was arbitrary we have shown (2.9) implies $p = S^*\mu$. The uniqueness of the adjoint operator proves the reverse implication.

For the proof of the stability estimate (2.10) see Theorem 2 in [Casas, 1985]. □

Remark 2.4. *We have assumed that the state equation is an elliptic PDE with Dirichlet boundary conditions. The theory in this chapter (and also the next chapter) can be adapted to elliptic PDEs with suitable Neumann boundary conditions, provided that $a(\cdot,\cdot)$ is still coercive. This is because the same regularity results hold for them and the same error estimates hold for their finite element approximations.*

2.2 Problem formulation

We are now in a position to formulate the optimal control problem precisely:

$$\begin{aligned} \min \quad & J(y,\eta) := \frac{1}{2}\sum_{\omega\in I}(y(\omega)-g_\omega)^2 + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)} \\ \text{over} \quad & C_0(\Omega)\times L^2(\Omega) \\ \text{s.t.} \quad & y = S\eta \text{ (i.e. (2.3) holds)} \\ \text{and} \quad & \eta \in U_{ad} := \{\eta \in L^2(\Omega) : a \le \eta \le b \text{ a.e. in } \Omega\}. \end{aligned} \tag{2.11}$$

Recall that $I \subset \Omega$ is a finite set of points, $\{g_\omega\}_{\omega\in I}$ are prescribed values at these points, and $\nu > 0$. We will prove results for the case that a and b are constant real numbers with $a < b$, and also the case of no control constraints (i.e. $b = -a = \infty$).

We can use the control-to-state operator S to define the reduced objective functional $\hat{J}(\eta) = J(S\eta, \eta)$. Then, as mentioned in Chapter 1, it is straightforward to show that (2.11) is equivalent to the optimisation problem:

$$\begin{aligned} \min \quad & \hat{J}(\eta) = \frac{1}{2}\sum_{\omega\in I}(S\eta(\omega)-g_\omega)^2 + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)} \\ \text{over} \quad & \eta \in U_{ad}. \end{aligned} \tag{2.12}$$

This equivalence is in the sense that $u \in U_{ad}$ solves (2.12) if and only if (Su, u) solves (2.11). It is simpler to work with the optimisation problem (2.12) for proving existence and uniqueness of a solution and deriving an optimality condition.

Theorem 2.5. *Problem (2.12) has a unique solution $u \in U_{ad}$, hence (2.11) has a unique solution (Su, u).*

Proof. This result follows using the same argument as is used for proving existence and uniqueness of solutions to standard optimal control problems, so we only outline the main ideas. See e.g. Theorem 2.14 in [Tröltzsch, 2010] for the details.

As $\hat{J} \ge 0$ we can construct an infimising sequence $\{\eta_n\} \subset U_{ad}$ i.e. a sequence such that $\hat{J}(\eta_n) \to \inf_{\eta\in U_{ad}} \hat{J}(\eta)$. Note that U_{ad} is a nonempty, closed, bounded and convex subset of a real reflexive Banach space, so it is weakly sequentially compact. This means there is a subsequence $\{\eta_{n_k}\}$ converging to some $u \in U_{ad}$. Since $S : L^2(\Omega) \to C_0(\Omega)$ is continuous, $\hat{J}$ is continuous. $\hat{J}$ is also convex, so it is weakly lower semicontinuous. Therefore u achieves the infimum of $\hat{J}$ i.e. it is a minimiser of $\hat{J}$. By a contradiction argument the strict convexity of $\hat{J}$ gives that u is the unique minimiser. □

Theorem 2.6. *$u \in U_{ad}$ is a solution of (2.12) if and only if there exists a $p \in L^2(\Omega)$ such that for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$, $p \in W_0^{1,q'}(\Omega)$ and*

$$u \in U_{ad}, \quad (p + \nu u, v - u) \geq 0 \qquad \forall v \in U_{ad}, \tag{2.13a}$$

$$a(v, p) = \sum_{\omega \in I} (Su(\omega) - g_\omega) v(\omega) \qquad \forall v \in W_0^{1,q}(\Omega). \tag{2.13b}$$

Proof. $\hat{J} : L^2(\Omega) \to \mathbb{R}$ has a Gâteaux derivative $J' : L^2(\Omega) \to L^2(\Omega)^*$. It is also (strictly) convex, and U_{ad} is a nonempty and convex subset of a real Banach space. So by a standard result (see e.g. Lemma 2.21 in [Tröltzsch, 2010]) $u \in L^2(\Omega)$ is a solution of (2.12) iff

$$u \in U_{ad}, \quad \langle \hat{J}'(u), v - u \rangle_{L^2(\Omega)^*, L^2(\Omega)} \geq 0 \quad \forall v \in U_{ad}. \tag{2.14}$$

For notational convenience define a function $g_d \in C^\infty(\bar{\Omega})$ such that $g_d(\omega) = g_\omega$ for all $\omega \in I$; such a function could be constructed using a mollifier. Let $\mu := \sum_{\omega \in I} \delta_\omega$, where δ_ω are Dirac measures centred at points ω, so $\mu \in \mathcal{M}(\Omega)$. Since $(Su - g_d)^2 \in C(\bar{\Omega})$ we can rewrite $\hat{J}$ as

$$\hat{J}(u) = \frac{1}{2} \int_\Omega (Su - g_d)^2 \mathrm{d}\mu + \frac{\nu}{2} \|u\|^2_{L^2(\Omega)}$$

and use the ideas of [Casas, 1986]. As a result our proof applies to objective functionals of this form with arbitrary $\mu \in \mathcal{M}(\Omega)$.

Calculating $\hat{J}'$ we find that (2.14) becomes

$$\int_\Omega (Su - g_d) S(v - u) \mathrm{d}\mu + \nu(u, v - u) \geq 0 \quad \forall v \in U_{ad}.$$

We now show that the first term on the left hand side can be written in the form $\int_\Omega p(v - u) \mathrm{d}x$, where p satisfies (2.13b).

For $u \in L^2(\Omega)$, $Su - g_d \in C(\bar{\Omega})$ and so it is measurable with respect to μ. So we can define a real Borel measure $\lambda_u : \mathcal{B} \to \mathbb{R}$ (where $\mathcal{B}$ denotes the Borel σ-algebra of Ω) by

$$\lambda_u(A) := \int_A (Su - g_d) \mathrm{d}\mu \quad \forall A \in \mathcal{B}. \tag{2.15}$$

Since μ is regular, we can check that λ_u is also regular. So λ_u is a real regular Borel measure (i.e. it belongs to $\mathcal{M}(\Omega)$) and Theorem 1.29 in [Rudin, 1987] says that for $z \in C_0(\Omega)$,

$$\int_\Omega (Su - g_d) z \mathrm{d}\mu = \int_\Omega z \mathrm{d}\lambda_u. \tag{2.16}$$

In particular, we can take $z := S(v-u)$ to get

$$\int_\Omega (Su - g_d)S(v-u)\mathrm{d}\mu = \int_\Omega S(v-u)\mathrm{d}\lambda_u = (S^*\lambda_u, v-u).$$

Let $p := S^*\lambda_u \in L^2(\Omega)$ then by (2.9), for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$, $p \in W_0^{1,q'}(\Omega)$ and

$$a(v,p) = \int_\Omega v\mathrm{d}\lambda_u \quad \forall v \in W_0^{1,q}(\Omega).$$

To finish, note that

$$\int_\Omega v\mathrm{d}\lambda_u = \int_\Omega (Su - g_d)v\mathrm{d}\mu = \sum_{\omega\in I}(Su(\omega) - g_\omega)v(\omega).$$

□

Corollary 2.7. *If $u \in U_{ad}$ is a solution of (2.12) then it has the additional regularity that $u \in W_0^{1,q'}(\Omega)$ for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$.*

Proof. Observe that (2.13a) is equivalent to

$$u(x) = \mathbb{P}_{[a,b]}\left(-\frac{1}{\nu}p(x)\right) \quad \text{for a.e. } x \in \Omega, \tag{2.17}$$

where $\mathbb{P}_{[a,b]}(v) := v + \max(0, a-v) - \max(0, v-b)$. If $v, w \in W_0^{1,q'}(\Omega)$ then $\max(v,w) \in W_0^{1,q'}(\Omega)$ (see for example [Morrey Jr., 1966]). So since $p \in W_0^{1,q'}(\Omega)$ for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$, we also get this additional regularity for u. □

2.2.1 Link to pointwise state constraints

We now discuss a link between the problem we consider in this chapter, which penalises deviation of the state from certain values at points, and an optimal control problem with a finite number of point state constraints i.e. a problem that forces the state to take certain values at points.

Consider the following problem, which is a generalisation of (2.11) in the case of no control constraints ($b = -a = \infty$):

$$\begin{aligned} \min \quad & J_\nu^\theta(y,\eta) := \frac{1}{2}\sum_{\omega\in I}(y(\omega) - g_\omega)^2 + \nu\left(\frac{1}{2}\theta\|y - g_d\|^2_{L^2(\Omega)} + \frac{1}{2}\|\eta\|^2_{L^2(\Omega)}\right) \\ \text{over} \quad & C_0(\Omega) \times L^2(\Omega) \\ \text{s.t.} \quad & (2.3) \text{ holds.} \end{aligned} \tag{2.18}$$

The modification is the addition of an optional $L^2(\Omega)$ fidelity term containing $g_d \in L^2(\Omega)$, which is weighted by $\theta \geq 0$. This allows us to relate (2.18) to a problem with point state constraints that is considered in the literature: In the limit $\nu \to 0$ we get convergence of solutions of (2.18) to the solution of the following problem, which can be found, for example, in [Leykekhman et al., 2013]:

$$\begin{aligned} \min \quad & J^\theta(y,\eta) := \frac{1}{2}\theta\|y - g_d\|^2_{L^2(\Omega)} + \frac{1}{2}\|\eta\|^2_{L^2(\Omega)} \\ \text{over} \quad & H^1_0(\Omega) \times L^2(\Omega) \\ \text{s.t.} \quad & (2.3) \text{ holds and } y(\omega) = g_\omega \text{ for } \omega \in I. \end{aligned} \tag{2.19}$$

Theorem 2.8. *Let (Su_ν, u_ν) solve (2.18) for $\nu > 0$ and $(S\bar{u}, \bar{u})$ solve (2.19). Then as $\nu \to 0$,*

$$\begin{aligned} Su_\nu &\rightharpoonup S\bar{u} \quad \text{in } H^1_0(\Omega) \\ u_\nu &\rightharpoonup \bar{u} \quad \text{in } L^2(\Omega). \end{aligned}$$

Proof. First note that there exists a function $\hat{u} \in L^2(\Omega)$ such that $S\hat{u}(\omega) = g_\omega$ for all $\omega \in I$ (see Lemma 1 in [Leykekhman et al., 2013]), so $J^\theta_\nu(Su_\nu, u_\nu) \leq \nu J^\theta(S\hat{u}, \hat{u})$. For all $\nu > 0$, $(S\hat{u}, \hat{u})$ is feasible for (2.18) so

$$\frac{\nu}{2}\|u_\nu\|^2_{L^2(\Omega)} \leq J^\theta_\nu(Su_\nu, u_\nu) \leq \nu J^\theta(S\hat{u}, \hat{u}) \leq C\nu \tag{2.20}$$

with C independent of ν. So u_ν is uniformly bounded with respect to ν in $L^2(\Omega)$, which means for every sequence $\nu_k \to 0$ there exists a sequence $u_{\nu_k} \rightharpoonup \tilde{u}$ in $L^2(\Omega)$. Moreover (2.20) and the stability result

$$\|Su_{\nu_k}\|_{H^1_0(\Omega)} \leq C\|u_{\nu_k}\|_{L^2(\Omega)}$$

with C independent of u_{ν_k} allows us to find a further subsequence, which we also denote by $\{\nu_k\}$, such that $Su_{\nu_k} \rightharpoonup \tilde{y}$ in $H^1_0(\Omega)$. Then taking the limit in (2.3) we see that $\tilde{y} = S\tilde{u}$. To complete the proof we need to show that $\tilde{u} = \bar{u}$, which we do by showing that $(S\tilde{u}, \tilde{u})$ is feasible for (2.19) and that $J^\theta(S\tilde{u}, \tilde{u}) \leq J^\theta(S\bar{u}, \bar{u})$.

Note that the same reasoning as for (2.20) gives $\frac{1}{\nu}\sum_{\omega\in I}(Su_\nu(\omega) - g_\omega)^2 \leq C$ independently of ν. Therefore we must have $Su_\nu(\omega) \to g_\omega$. So $S\tilde{u}(\omega) = g_\omega$ for $\omega \in I$ and $(S\tilde{u}, \tilde{u})$ is feasible for (2.19).

The weak lower semicontinuity of J^θ and $J^\theta(Su_{\nu_k}, u_{\nu_k}) \leq \frac{J^\theta_{\nu_k}(Su_{\nu_k}, u_{\nu_k})}{\nu_k}$ im-

plies

$$J^\theta(S\tilde{u},\tilde{u}) \le \liminf_{k\to\infty} J^\theta(Su_{\nu_k},u_{\nu_k}) \le \liminf_{k\to\infty} \frac{J^\theta_{\nu_k}(Su_{\nu_k},u_{\nu_k})}{\nu_k}.$$

Also the optimality of (Su_{ν_k},u_{ν_k}) for (2.18) and $\frac{J^\theta_{\nu_k}(S\bar{u},\bar{u})}{\nu_k} = J^\theta(S\bar{u},\bar{u})$ implies

$$\liminf_{k\to\infty} \frac{J^\theta_{\nu_k}(Su_{\nu_k},u_{\nu_k})}{\nu_k} \le \liminf_{k\to\infty} \frac{J^\theta_{\nu_k}(S\bar{u},\bar{u})}{\nu_k} = J^\theta(S\bar{u},\bar{u}).$$

Combining these we get

$$J^\theta(S\tilde{u},\tilde{u}) \le J^\theta(S\bar{u},\bar{u}),$$

so we have proved the result. □

2.3 Discretisation

In this section we discretise the state equation using a finite element method and use this to formulate two different discrete problems. We then derive discrete optimality conditions for each problem.

We now make some slightly stronger assumptions on Ω and A than were necessary for the problem formulation and analysis in the previous section. From now onwards we assume that Ω is convex. This simplifies the presentation since then the finite element space for the state (defined shortly) is a subset of $C_0(\Omega)$. Note that if the state equation had Neumann boundary conditions (see Remark 2.4) then nonconvex domains would not cause this complication. Also from now onwards assume that the boundary of Ω and the coefficient functions a_{ij} and a_0 in the elliptic operator A are sufficiently smooth that for $2 \le s < \frac{2n}{n-2}$,

$$\|S\eta\|_{W^{2,s}(\Omega)} \le C\|\eta\|_{L^s(\Omega)} \quad \forall \eta \in L^s(\Omega). \tag{2.21}$$

This holds, for example, when $A = -\Delta$ and Ω is smooth (see e.g. Theorem 9.9 in [Gilbarg and Trudinger, 2001]).

Since Ω is convex with a sufficiently smooth boundary, we can take a family of polygonal or polyhedral approximations $\Omega_h \subset \Omega$ such that the vertices of $\partial\Omega_h$ lie on $\partial\Omega$ and $|\Omega \setminus \Omega_h| \le Ch^2$. On each Ω_h we can construct a conforming triangulation T_h of triangles or tetrahedra T with maximum diameter $h := \max_{T\in T_h} h(T)$, where $h(T)$ is the diameter of an element T. Additionally suppose that the family of triangulations are conforming and quasi-uniform i.e. there exists a constant C such

that

$$\frac{h(T)}{\rho(T)} \leq C \quad \forall T \in T_h,$$

where $\rho(T)$ is the radius of the largest ball contained in T, and there exists a constant C such that

$$\frac{h}{h(T)} \leq C \quad \forall T \in T_h$$

(see e.g. Chapter 3 of [Ciarlet, 1978]). We can define the following family of discrete spaces of piecewise linear globally continuous finite elements which vanish on the boundary:

$$V_h := \{v_h \in C_0(\Omega) : v_h|_T \in P_1(T) \text{ for all } T \in T_h \text{ and } v_h|_{\Omega \setminus \Omega_h} = 0\}.$$

Here $P_1(T)$ is the set of affine functions over T. Our motivation for using this finite element space (rather than, for example, a space of piecewise constant finite elements) is that it is a subspace of $C_0(\Omega)$.

We also construct a family of triangulations T^σ of triangles or tetrahedra with maximum element diameter σ. We allow elements on the boundary to have one curved face, and assume that T^σ is conforming and shape regular (as we did for T_h). Note that the family of triangulations T^σ potentially has nothing in common with T_h. We can now define the following discrete space $U_{ad,\sigma}$ for the control:

$$\begin{aligned} U_\sigma &:= \{u_\sigma \in C(\bar{\Omega}) : u_\sigma|_T \in P_1(T) \text{ for all } T \in T^\sigma\}, \\ U_{ad,\sigma} &:= \{u_\sigma \in U_\sigma : a \leq u_\sigma \leq b\}. \end{aligned}$$

This is a space of piecewise linear globally continuous finite elements (as was V_h) with $U_{ad,\sigma} \subset U_{ad}$, however we do not require the functions to vanish at the boundary. Recall from Corollary 2.7 that $u \in W_0^{1,q'}(\Omega)$ for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$, and piecewise constant finite elements approximate such functions almost as well as piecewise linear finite elements. However we define $U_{ad,\sigma}$ to use piecewise linear finite elements as we want to allow taking the same discrete space for the control and state. This can simplify implementations.

For U_σ the following approximation property holds: There exists an interpolation operator $\Pi_\sigma : W^{l,p}(\Omega) \to U_\sigma$ $(l = 1, 2;\ 1 \leq p < \infty)$ such that

$$\|v - \Pi_\sigma v\|_{W^{m,p}(\Omega)} \leq C\sigma^{1-m}\|v\|_{W^{1,p}(\Omega)}, \quad m = 0, 1. \tag{2.22}$$

Such an interpolation operator can be defined as in [Scott and Zhang, 1990]. It also has the property that $v \in U_{ad}$ implies $\Pi_\sigma v \in U_{ad}$.

We now introduce discrete approximations of S and S^*. Define $S_h : L^2(\Omega) \to C_0(\Omega)$ by $S_h\eta = y_h$, where y_h satisfies

$$y_h \in V_h, \quad a(y_h, v_h) = (\eta, v_h) \quad \forall v_h \in V_h. \tag{2.23}$$

It is a standard result that this problem has a unique solution.

We now prove some supremum norm error estimates for S_h that will be useful for the numerical analysis.

Lemma 2.9. *For $\eta \in L^s(\Omega)$ and $2 \leq s < \infty$,*

$$\|S\eta - S_h\eta\|_\infty \leq C(s)h^{2-\frac{n}{s}}\|\eta\|_{L^s(\Omega)}, \quad n = 2, 3. \tag{2.24}$$

Proof. First we will recall some results from the literature that hold under the assumptions we make in this paper. By (34) in [Leykekhman et al., 2013] we have that

$$\|Sv - S_hv\|_{L^s(\Omega)} \leq C(s)h^2\|Sv\|_{W^{2,s}(\Omega)} \quad \forall v \in L^s(\Omega).$$

This was originally proved for $n = 2$ on p438 in [Rannacher and Scott, 1982]. Applying an inverse inequality on each element of the triangulation gives that

$$\|v_h\|_{L^\infty(\Omega_h)} \leq C(s)h^{-\frac{n}{s}}\|v_h\|_{L^s(\Omega)} \quad \forall v_h \in V_h \tag{2.25}$$

(see e.g. [Ciarlet, 1978]). Similarly, for the piecewise linear interpolation operator $I_h : C_0(\Omega) \to V_h$ and $r \in [1, \infty]$ we have

$$\|v - I_hv\|_{L^r(\Omega_h)} \leq C(s)h^{2+\frac{1}{r}-\frac{1}{s}}\|v\|_{W^{2,s}(\Omega_h)} \quad \forall v \in W^{2,s}(\Omega).$$

(see e.g. Theorem 3.1.5 in [Ciarlet, 1978]).

Combining these results we get that

$$\begin{aligned}
\|S\eta - S_h\eta\|_{L^\infty(\Omega_h)} &\leq \|S\eta - I_hS\eta\|_{L^\infty(\Omega_h)} + \|I_hS\eta - S_h\eta\|_{L^\infty(\Omega_h)}, \\
&\leq C(s)(h^{2-\frac{n}{s}}\|S\eta\|_{W^{2,s}(\Omega)} + h^{-\frac{n}{s}}\|I_hS\eta - S_h\eta\|_{L^s(\Omega_h)}) \\
&\leq C(s)h^{-\frac{n}{s}}(h^2\|S\eta\|_{W^{2,s}(\Omega)} + \|I_hS\eta - S\eta\|_{L^s(\Omega_h)} + \|S\eta - S_h\eta\|_{L^s(\Omega_h)}) \\
&\leq C(s)h^{2-\frac{n}{s}}(\|S\eta\|_{W^{2,s}(\Omega)} + \|\eta\|_{L^s(\Omega)}) \\
&\leq C(s)h^{2-\frac{n}{s}}\|\eta\|_{L^s(\Omega)}.
\end{aligned}$$

We now need to prove a supremum norm error estimate for the skin $\Omega \setminus \Omega_h$. By Theorem 4.12 Part II in [Adams and Fournier, 2003]:

- If $s \geq n$ then $W^{2,s}(\Omega) \hookrightarrow C^{0,\lambda}(\bar{\Omega})$ for $0 < \lambda < 1$.
- If $\frac{n}{2} < s < n$ then $W^{2,s}(\Omega) \hookrightarrow C^{0,\lambda}(\bar{\Omega})$ for $0 < \lambda \leq 2 - \frac{n}{s}$.

Let

$$\bar{\lambda}(s) := \begin{cases} 1 - \frac{n}{2s} & s \geq n, \\ 2 - \frac{n}{s} & \frac{n}{2} < s < n, \end{cases}$$

then for $x_1 \in \Omega \setminus \Omega_h$ we have

$$\inf_{x_2 \in \partial\Omega} |S\eta(x_1) - S\eta(x_2)| \leq C(s) \inf_{x_2 \in \partial\Omega} |x_1 - x_2|^{\bar{\lambda}(s)}.$$

Since Ω has a sufficiently smooth boundary our construction of Ω_h means that

$$\inf_{x_2 \in \partial\Omega} |x_1 - x_2| \leq Ch^2 \quad \forall x_1 \in \Omega \setminus \Omega_h.$$

Note that for our range of s we have $h^{2\bar{\lambda}(s)} \leq Ch^{2-\frac{n}{s}}$ for sufficiently small h. Using this and $S\eta|_{\partial\Omega} = 0$ we get

$$|S\eta(x_1)| \leq C(s)h^{2-\frac{n}{s}} \quad \forall x_1 \in \Omega \setminus \Omega_h.$$

Hence

$$\|S\eta - S_h\eta\|_\infty \leq \max(\|S\eta - S_h\eta\|_{L^\infty(\Omega_h)}, \|S\eta - S_h\eta\|_{L^\infty(\Omega\setminus\Omega_h)}) \leq C(s)h^{2-\frac{n}{s}}\|\eta\|_{L^s(\Omega)}.$$

□

Corollary 2.10. *For $\eta \in L^2(\Omega)$,*

$$\|S\eta - S_h\eta\|_\infty \leq Ch^{2-\frac{n}{2}}\|\eta\|_{L^2(\Omega)}, \quad n = 2, 3. \tag{2.26}$$

For $\eta \in W^{1,q'}(\Omega)$ with $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$,

$$\|S\eta - S_h\eta\|_\infty \leq C(q')h^{3-\frac{n}{q'}}\|\eta\|_{W^{1,q'}(\Omega)} \quad n = 2, 3. \tag{2.27}$$

For $\eta \in H^1(\Omega)$ and any $\varepsilon > 0$,

$$\|S\eta - S_h\eta\|_\infty \leq \|\eta\|_{H^1(\Omega)} \begin{cases} C(\varepsilon)h^{2-\varepsilon} & n = 2, \\ Ch^{\frac{3}{2}} & n = 3. \end{cases} \tag{2.28}$$

Proof. The first estimate follows by taking $s = 2$ in Lemma 2.9. The other estimates follow by combining the lemma with Sobolev embedding results. In particular, if

$\eta \in W_0^{1,q'}(\Omega)$ with $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$ then $W_0^{1,q'}(\Omega) \hookrightarrow L^s(\Omega)$ for $s = \frac{nq'}{n-q'} \geq 2$. So

$$C(s)h^{2-\frac{n}{s}}\|\eta\|_{L^s(\Omega)} \leq C(q')h^{3-\frac{n}{q'}}\|\eta\|_{W_0^{1,q'}(\Omega)},$$

which proves the second estimate. Note that this estimate is proved in a similar way in Theorem 3 in [Leykekhman et al., 2013]. If $\eta \in H^1(\Omega)$ then

$$H^1(\Omega) \hookrightarrow L^s(\Omega) \quad \forall s \in \begin{cases} [1,\infty) & n=2, \\ [1,6] & n=3. \end{cases}$$

So by taking s sufficiently large for $n = 2$ and $s = 6$ for $n = 3$ we get: For any $\varepsilon > 0$,

$$C(s)h^{2-\frac{n}{s}}\|\eta\|_{L^s(\Omega)} \leq \begin{cases} C(\varepsilon)h^{2-\varepsilon}\|\eta\|_{H^1(\Omega)} & n=2, \\ Ch^{\frac{3}{2}}\|\eta\|_{H^1(\Omega)} & n=3. \end{cases}$$

This proves the third estimate. □

We will use (2.26) in Section 2.4.1 and (2.27) in Section 2.4.2 to prove $L^2(\Omega)$ error estimates for the point optimal control problem. We will use (2.28) in Chapter 3 to prove an $L^2(\Omega)$ error estimate for a different optimal control problem.

Since S_h is continuous (which follows from (2.26)) and linear it has an adjoint operator $S_h^* : \mathcal{M}(\Omega) \to L^2(\Omega)$. Note that the same calculation as in Lemma 2.3 gives that $p_h = S_h^*\mu$ if and only if p_h satisfies

$$p_h \in V_h, \quad a(v_h, p_h) = \int_\Omega v_h \mathrm{d}\mu \quad \forall v_h \in V_h. \tag{2.29}$$

We have the following error estimate for S_h^*, which we will use in Section 2.4.1: For $\mu \in \mathcal{M}(\Omega)$,

$$\|S^*\mu - S_h^*\mu\|_{L^2(\Omega)} \leq Ch^{2-\frac{n}{2}}\|\mu\|_{\mathcal{M}(\Omega)}, \tag{2.30}$$

with C independent of μ and h. This follows by noting that for any $v \in L^2(\Omega)$,

$$(S^*\mu - S_h^*\mu, v) = \langle \mu, Sv - S_hv\rangle_{\mathcal{M}(\Omega)} \leq \|\mu\|_{\mathcal{M}(\Omega)}\|Sv - S_hv\|_\infty.$$

Then using (2.26) gives the result. This was proved in Theorem 3 in [Casas, 1985] for convex polygonal domains, but the result also holds for domains with sufficiently smooth boundaries because (2.26) does. Related theory was developed in [Scott, 1973].

Remark 2.11. *The estimates in Lemma 2.9 and Corollary 2.10 still hold if S and*

S_h are appropriately defined control-to-state operators corresponding to an elliptic PDE with Neumann boundary conditions.

2.3.1 Discrete problems

We are now ready to introduce the two discrete problems that we consider in our numerical analysis.

Define the discrete reduced objective functional $\hat{J}_h : L^2(\Omega) \to \mathbb{R}$ by

$$\hat{J}_h(\eta) = J(S_h\eta, \eta) = \frac{1}{2}\sum_{\omega\in I}(S_h\eta(\omega) - g_\omega)^2 + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)}.$$

Then the first discrete problem we consider is ($\mathrm{M1}_h$):

$$\min \hat{J}_h(\eta_\sigma) \text{ over } \eta_\sigma \in U_{ad,\sigma}. \tag{2.31}$$

Proposition 2.12. *There is a unique solution $u_{\sigma,h} \in U_{ad,\sigma}$ to ($\mathrm{M1}_h$) (see (2.31)). Moreover, $u_{\sigma,h} \in U_{ad,\sigma}$ is a solution of ($\mathrm{M1}_h$) if and only if there exists $p_h \in V_h$ such that*

$$u_{\sigma,h} \in U_{ad,\sigma}, \quad (p_h + \nu u_{\sigma,h}, v_\sigma - u_{\sigma,h}) \geq 0 \qquad \forall v_\sigma \in U_{ad,\sigma} \tag{2.32a}$$

$$a(v_h, p_h) = \sum_{\omega\in I}(S_h u_{\sigma,h}(\omega) - g_\omega)v_h(\omega) \qquad \forall v_h \in V_h. \tag{2.32b}$$

Proof. The proof follows from the same considerations as in Theorems 2.5 and 2.6. Note that $p_h = S_h^*\lambda_{h,u_{\sigma,h}}$ where for $\eta \in L^2(\Omega)$ we define $\lambda_{h,\eta} \in \mathcal{M}(\Omega)$ by

$$\lambda_{h,\eta}(A) = \int_A (S_h\eta - g_d)\mathrm{d}\mu \quad \forall A \in \mathcal{B} \tag{2.33}$$

with $\mu = \sum_{\omega\in I}\delta_\omega$. □

We refer to ($\mathrm{M1}_h$) as the explicitly discretised problem as we make the control belong to a space of discrete functions.

Alternatively we could use the variational discretisation concept from [Hinze, 2005] and leave the control in the infinite dimensional space U_{ad}. This leads to the potentially different (see Remark 2.14) discrete problem ($\mathrm{M2}_h$):

$$\min \hat{J}_h(\eta) \text{ over } \eta \in U_{ad}. \tag{2.34}$$

Proposition 2.13. *There is a unique solution $u_h \in U_{ad}$ to ($\mathrm{M2}_h$) (see (2.34)). Moreover, $u_h \in L^2(\Omega)$ is a solution of ($\mathrm{M2}_h$) if and only if there exists $p_h \in V_h$ such*

that

$$u_h \in U_{ad}, \quad (p_h + \nu u_h, v - u_h) \geq 0 \qquad \forall v \in U_{ad} \tag{2.35a}$$

$$a(v_h, p_h) = \sum_{\omega \in I} (S_h u_h(\omega) - g_\omega) v_h(\omega) \qquad \forall v_h \in V_h. \tag{2.35b}$$

Proof. The proof also follows from the same considerations as in Theorems 2.5 and 2.6. □

A priori we only know that u_h belongs to U_{ad}. However observe that (2.35a) can be expressed using the pointwise projection operator $\mathbb{P}_{[a,b]}$ from (2.17) as

$$u_h = \mathbb{P}_{[a,b]}\Big(-\frac{1}{\nu} p_h \Big).$$

So (2.35a) has a simpler form than (2.32a), which is an $L^2(\Omega)$ projection onto a discrete space. This means u_h inherits a piecewise linear structure from $p_h \in V_h$, but observe that u_h does not necessarily belong to V_h due to the control constraints. We refer to this as an implicit discretisation; we are not requiring u_h to be a piecewise linear function, but it gains this property indirectly through the discretisation of the state. Even though u_h does not necessarily belong to V_h, this problem can be solved computationally. We will elaborate on this in Section 2.5.1.

Remark 2.14. *The motivation for the implicitly discretised problem* ($\mathrm{M2}_h$) *is that it allows a better approximation of the set where the control constraints are active (indicated in Figure 2.1), likely leading to a smaller error. For a more thorough explanation see [Hinze, 2005].*

Remark 2.15. *Note that if there are no active control constraints (e.g. if* $b = -a = \infty$*) and* $V_h \subset U_\sigma$*, then* ($\mathrm{M1}_h$) *and* ($\mathrm{M2}_h$) *are equivalent. In order for* $V_h \subset U_\sigma$ *we need "*$T^\sigma \subset T_h$*". By this we mean that each element of* T^σ *is contained in either a single element of* T_h *or the skin* $\Omega \setminus \Omega_h$*.*

2.4 Numerical analysis

We now prove $L^2(\Omega)$ error estimates between the solution of the continuous problem (2.12) and the two discrete problems ($\mathrm{M1}_h$) and ($\mathrm{M2}_h$) (see (2.31) and (2.34)). We use two different approaches for this numerical analysis. Approach 1 in the next section allows us to prove error estimates for the two discrete problems in most (but not all) the cases we would like. Approach 2 in Section 2.4.2 only reproduces some

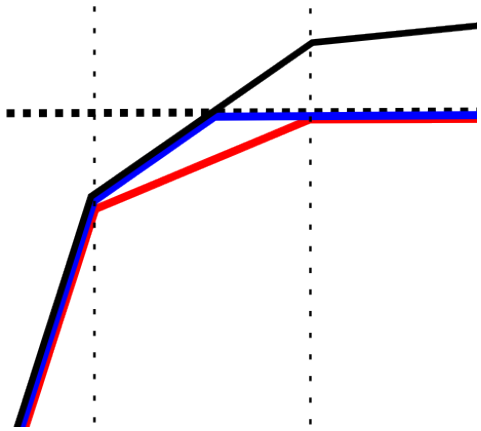


Figure 2.1: An illustration in 1D of how u_h is determined by p_h (black line) when the discrete space for the control and state are the same and $\nu = 1$. The horizontal dashed line is the value of b and the vertical dashed lines marks the boundary between elements. The blue line is u_h calculated from p_h using (2.32a) and the red line is using (2.35a). Assuming the p_h are similar and good approximations of p for both ($\mathrm{M1}_h$) and ($\mathrm{M2}_h$), this suggests that ($\mathrm{M2}_h$) will give a better approximation of u.

of these error estimates. However it is simpler, which will allow us to adapt it to a more complicated setting in the next chapter, where we prove error estimates for an optimal control problem involving surfaces of codimension 1.

2.4.1 Approach 1

This error analysis is based on [Casas and Tröltzsch, 2003], where an a priori $L^2(\Omega)$ error estimate is proved for the standard optimal control problem that we introduced in Chapter 1. The approach allows us to prove $L^2(\Omega)$ error estimates for both ($\mathrm{M1}_h$) and ($\mathrm{M2}_h$). The only estimates it does not give are ones for ($\mathrm{M1}_h$) when $n = 3$ (but we are not able to prove these using Approach 2 either). In particular we will get the following results.

Theorem 2.16. *Assume $n = 2$. Let u solve (2.12) and $u_{\sigma,h}$ solve ($\mathrm{M1}_h$) (see (2.31)). Then*

$$\|u - u_{\sigma,h}\|_{L^2(\Omega)} \leq C(\sqrt{\sigma} + h)$$

with C independent of σ and h.

Theorem 2.17. *Assume $n = 2$ or 3. Let u solve (2.12) and u_h solve ($\mathrm{M2}_h$) (see (2.34)). Then*

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{2-\frac{n}{2}}$$

with C independent of σ and h.

Corollary 2.18. *Assume $n = 2$ or 3, there are no active control constraints (e.g. $b = -a = \infty$), and that $V_h \subset U_\sigma$. Let u solve (2.12) and $u_{\sigma,h}$ solve* ($\mathrm{M1}_h$) *(see (2.31)). Then*

$$\|u - u_{\sigma,h}\|_{L^2(\Omega)} \leq Ch^{2-\frac{n}{2}}$$

with C independent of σ and h.

Proof. This result follows from the equivalence between ($\mathrm{M1}_h$) and ($\mathrm{M2}_h$) that is highlighted in Remark 2.15. □

Note that these results suggest ($\mathrm{M2}_h$) is the preferred discretisation. In particular, we can only prove an error estimate in the case of $n = 3$ with control constraints for ($\mathrm{M2}_h$). Also the error estimate in the case of $n = 2$ with control constraints is better for ($\mathrm{M2}_h$).

The idea of the approach is to consider the error caused by the discretisation of the control and state separately, then combine them. This approach only needs the weak supremum norm error estimate for the state equation (where as a stronger one is needed for Approach 2 in Section 2.4.2), but it does require an error estimate for the adjoint of the control-to-state operator. An advantage of this approach is that it can give insight into the best choice of triangulations for the control and state, which are not necessarily the same.

To begin we split the error as

$$\|u - u_{\sigma,h}\|_{L^2(\Omega)} \leq \|u - u_\sigma\|_{L^2(\Omega)} + \|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)}, \tag{2.36}$$

where u_σ solves the semi discrete control problem

$$\min \hat{J}(\eta_\sigma) \text{ over } \eta_\sigma \in U_{ad,\sigma}. \tag{2.37}$$

Proposition 2.19. *There is a unique solution $u_\sigma \in U_{ad,\sigma}$ to (2.37). Moreover, $u_\sigma \in U_{ad,\sigma}$ is a solution of (2.37) if and only if there exists a $p_\sigma \in L^2(\Omega)$ such that for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$, $p_\sigma \in W_0^{1,q'}(\Omega)$ and*

$$u \in U_{ad,\sigma} \quad (p_\sigma + \nu u_\sigma, v_\sigma - u_\sigma) \geq 0 \qquad \forall v_\sigma \in U_{ad,\sigma} \tag{2.38a}$$

$$a(v, p_\sigma) = \sum_{\omega \in I} (Su_\sigma(\omega) - g_\omega)v(\omega) \qquad \forall v \in W_0^{1,q}(\Omega). \tag{2.38b}$$

Note that p_σ is not a discrete function. The subscript σ is to denote association with the discrete control u_σ.

Proof. $U_{ad,\sigma}$ is still a closed convex subset of $L^2(\Omega)$ so the proofs in Theorems 2.5 and 2.6 apply. Note that $p_\sigma = S^*\lambda_{u_\sigma}$ where $\lambda_{u_\sigma} \in \mathcal{M}(\Omega)$ is defined analogously to (2.15) by

$$\lambda_{u_\sigma}(A) := \int_A (Su_\sigma - g_d)\mathrm{d}\mu \quad \forall A \in \mathcal{B}. \tag{2.39}$$

□

Whereas (2.34) minimises the discrete reduced objective functional over the continuous space, this problem minimises the continuous reduced objective functional over the discrete space. So the solution of (2.37) is discrete, but the corresponding state is continuous, and this problem cannot be solved computationally.

The first term on the right hand side of (2.36) can be thought of as the error from the discretisation of the control, as we are comparing the minimiser of the continuous objective functional over continuous and discrete controls. Similarly the second term on the right hand side of (2.36) can be thought of as the error from the discretisation of the state, as we compare the minimiser of the continuous and discrete objective functionals, both over discrete controls. To prove Theorem 2.16 it is sufficient to prove an error estimate for each term separately, which we do in Lemmas 2.21 and 2.22. Note that we have additional assumptions in Theorem 2.16 because we need these in order to prove Lemma 2.21. But first we will prove some a priori estimates for the solution of (2.37).

Lemma 2.20. *Let u_σ solve (2.37) and p_σ satisfy the optimality system (2.38). For all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$,*

$$\|u_\sigma\|_{L^2(\Omega)} + \|Su_\sigma\|_{L^2(\Omega)} + \|p_\sigma\|_{W_0^{1,q'}(\Omega)} \leq C(q') \tag{2.40}$$

with C independent of σ. Moreover, when $n = 2$ there exists some $q > n$ such that

$$\|u_\sigma\|_{L^q(\Omega)} + \|p_\sigma\|_{L^q(\Omega)} \leq C \tag{2.41}$$

with C independent of σ.

Proof. Using (2.10), (2.39) and (2.4), for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$ we have

$$\begin{aligned}\|p_\sigma\|_{W_0^{1,q'}(\Omega)} &\leq C(q')\|\lambda_{u_\sigma}\|_{\mathcal{M}(\Omega)} \\ &= C(q')\sum_{\omega\in I}|Su_\sigma - g_d| \\ &\leq C(q')(\|Su_\sigma\|_\infty + \max_{\omega\in I}|g_\omega|) \\ &\leq C(q')(\|u_\sigma\|_{L^2(\Omega)} + 1). \end{aligned} \tag{2.42}$$

Combining this with (2.4) again we get

$$\|u_\sigma\|_{L^2(\Omega)} + \|Su_\sigma\|_{L^2(\Omega)} + \|p_\sigma\|_{W_0^{1,q'}(\Omega)} \leq C(q')(\|u_\sigma\|_{L^2(\Omega)} + 1). \tag{2.43}$$

If $a, b \in \mathbb{R}$ then

$$\|u_\sigma\|_{L^2(\Omega)} \leq |\Omega|^{\frac{1}{2}} \max(|a|, |b|).$$

If $b = -a = \infty$ then $0 \in U_{ad,\sigma}$, so $\hat{J}(u_\sigma) \leq \hat{J}(0)$. Since $S0 = 0$, this means

$$\frac{\nu}{2}\|u_\sigma\|^2_{L^2(\Omega)} \leq \frac{1}{2}\sum_{\omega\in I} g_\omega^2. \tag{2.44}$$

So regardless of the assumptions on a and b, we have $\|u_\sigma\|_{L^2(\Omega)} \leq C$. Combining this with (2.43) gives the first bound (2.40).

For the second bound we assume $n = 2$. If $a, b \in \mathbb{R}$ then we have

$$\|u_\sigma\|_{L^q(\Omega)} \leq |\Omega|^{\frac{1}{q}} \max(|a|, |b|).$$

If $b = -a = \infty$ we can use the $L^q(\Omega)$ stability of the $L^2(\Omega)$ projection (see e.g. [Crouzeix and Thomée, 1987]) to get $\|u_\sigma\|_{L^q(\Omega)} \leq \frac{1}{\nu}\|p_\sigma\|_{L^q(\Omega)}$. So for all $q > 2$,

$$\|u_\sigma\|_{L^q(\Omega)} + \|p_\sigma\|_{L^q(\Omega)} \leq C(\|p_\sigma\|_{L^q(\Omega)} + 1).$$

We now need some $q > 2$ such that $\|p_\sigma\|_{L^q(\Omega)} \leq C$ independently of σ. By Sobolev embedding results, if $s > \frac{n}{2}$ then $W_0^{1,s}(\Omega) \hookrightarrow L^t(\Omega)$ for some $t > n$. In particular for $n = 2$ we can take $s = \frac{4}{3} > \frac{n}{2} = 1$, since $p_\sigma \in W_0^{1,s}(\Omega)$ for $s \in (\frac{2n}{n+2}, \frac{n}{n-1}) = (1, 2)$. Then for some $q > 2$,

$$\|p_\sigma\|_{L^q(\Omega)} \leq C\|p_\sigma\|_{W_0^{1,\frac{4}{3}}(\Omega)} \leq C,$$

where we have used (2.40) for the final inequality. Note that for $n = 3$ we would

require $s > \frac{n}{2} = \frac{3}{2}$, but for such an s we do not have $p_\sigma \in W_0^{1,s}(\Omega)$, which only holds when $s \in (\frac{2n}{n+2}, \frac{n}{n-1}) = (\frac{5}{4}, \frac{3}{2})$. □

Lemma 2.21 (Error from discretisation of the control)**.** *Assume $n = 2$. Let u and u_σ be solutions of (2.12) and (2.37) respectively. Then*

$$\|u - u_\sigma\|_{L^2(\Omega)} \leq C\sqrt{\sigma}$$

with C independent of σ (and h).

Proof. Test with $v = u_\sigma$ in (2.13a) to get

$$(p + \nu u, u_\sigma - u) \geq 0.$$

Test with $v_\sigma = \Pi_\sigma u$ in (2.38a) to get

$$(p_\sigma + \nu u_\sigma, \Pi_\sigma u - u_\sigma) = (p_\sigma + \nu u_\sigma, \Pi_\sigma u - u) + (p_\sigma + \nu u_\sigma, u - u_\sigma) \geq 0.$$

Adding these two inequalities and rearranging we get

$$\nu\|u - u_\sigma\|^2_{L^2(\Omega)} + (p_\sigma - p, u_\sigma - u) \leq (p_\sigma + \nu u_\sigma, \Pi_\sigma u - u). \tag{2.45}$$

Recall from the proof of Theorem 2.6 that $p = S^*\lambda_u$ with λ_u defined by (2.15). Similarly during the proof of Proposition 2.19 we find that $p_\sigma = S^*\lambda_{u_\sigma}$ with λ_{u_σ} defined by (2.39). So using this and Theorem 1.29 in [Rudin, 1987] (see e.g. (2.16)) we get

$$\begin{aligned}(p_\sigma - p, u_\sigma - u) =& (S^*\lambda_{u_\sigma} - S^*\lambda_u, u_\sigma - u) = \langle \lambda_{u_\sigma} - \lambda_u, S(u_\sigma - u)\rangle_{\mathcal{M}(\Omega)} \\ &= \int_\Omega (S(u - u_\sigma))^2 \mathrm{d}\mu \geq 0.\end{aligned}$$

This means the second term on the left hand side of (2.45) can be dropped.

We now bound the right hand side of (2.45). By Lemma 2.20, for $n = 2$ there exists some $q > n$ such that $\|u_\sigma\|_{L^q(\Omega)}$ and $\|p_\sigma\|_{L^q(\Omega)}$ are bounded independently of σ. So using Hölder's inequality with this q we get

$$\begin{aligned}(p_\sigma + \nu u_\sigma, \Pi_\sigma u - u) &\leq \|p_\sigma + \nu u_\sigma\|_{L^q(\Omega)} \|\Pi_\sigma u - u\|_{L^{q'}(\Omega)} \\ &\leq (\|p_\sigma\|_{L^q(\Omega)} + \nu\|u_\sigma\|_{L^q(\Omega)})\|\Pi_\sigma u - u\|_{L^{q'}(\Omega)} \\ &\leq C\|\Pi_\sigma u - u\|_{L^{q'}(\Omega)},\end{aligned}$$

with C independent of σ. Now (2.22) gives

$$\|\Pi_\sigma u - u\|_{L^{q'}(\Omega)} \leq C\sigma \|u\|_{W_0^{1,q'}(\Omega)} \leq C\sigma,$$

so we can deduce that

$$(p_\sigma + \nu u_\sigma, \Pi_\sigma u - u) \leq C\sigma.$$

Recall from Lemma 2.20 that $\|u_\sigma\|_{L^q(\Omega)}$ and $\|p_\sigma\|_{L^q(\Omega)}$ are not bounded independently of σ for $n = 3$, so the above proof does not work in that case.

□

Lemma 2.22 (Error from discretisation of the state)**.** *Assume $n = 2$ or 3. Let u_σ and $u_{\sigma,h}$ be the solutions of (2.37) and* ($\mathrm{M1}_h$) *(see (2.31)) respectively. Then*

$$\|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \leq Ch^{2-\frac{n}{2}}$$

with C independent of σ and h.

Proof. Testing (2.32a) with $v_\sigma = u_\sigma$ gives

$$(p_h + \nu u_{\sigma,h}, u_\sigma - u_{\sigma,h}) \geq 0.$$

Testing (2.38a) with $v_h = u_{\sigma,h}$ gives

$$(p_\sigma + \nu u_\sigma, u_{\sigma,h} - u_\sigma) \geq 0.$$

Adding these two inequalities, using that $p_h = S_h^* \lambda_{h,u_{\sigma,h}}$ and $p_\sigma = S^* \lambda_{u_\sigma}$, and introducing $S_h^* \lambda_{h,u_\sigma}$ (see (2.33)) we get

$$\begin{aligned}
\nu \|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)}^2 \leq & (p_h - p_\sigma, u_\sigma - u_{\sigma,h}) \\
= & (S_h^* \lambda_{u_{\sigma,h}} - S^* \lambda_{u_\sigma}, u_\sigma - u_{\sigma,h}) \\
\leq & (S_h^* \lambda_{h,u_{\sigma,h}} - S_h^* \lambda_{h,u_\sigma}, u_\sigma - u_{\sigma,h}) \\
& + (S_h^* \lambda_{h,u_\sigma} - S^* \lambda_{u_\sigma}, u_\sigma - u_{\sigma,h}).
\end{aligned} \tag{2.46}$$

Note that

$$\begin{aligned}
(S_h^* \lambda_{h,u_{\sigma,h}} - S_h^* \lambda_{h,u_\sigma}, u_\sigma - u_{\sigma,h}) = & \langle \lambda_{h,u_{\sigma,h}} - \lambda_{h,u_\sigma}, S_h(u_\sigma - u_{\sigma,h}) \rangle_{\mathcal{M}(\Omega)} \\
= & - \int_\Omega (S_h(u_{\sigma,h} - u_\sigma))^2 \mathrm{d}\mu \leq 0.
\end{aligned}$$

So the first term on the right hand side of (2.46) can be dropped. Also note that

$$(S_h^*\lambda_{h,u_\sigma} - S^*\lambda_{u_\sigma}, u_\sigma - u_{\sigma,h}) = (S_h^*\lambda_{h,u_\sigma} - S^*\lambda_{h,u_\sigma}, u_\sigma - u_{\sigma,h}) + (S^*\lambda_{h,u_\sigma} - S^*\lambda_{u_\sigma}, u_\sigma - u_{\sigma,h}),$$

and we can bound both term on the right hand side of this. Using (2.30) and

$$\|\lambda_{h,u_\sigma}\|_{\mathcal{M}(\Omega)} = \sum_{\omega\in I} |S_h u_\sigma(\omega) - g_d| \le C\|S_h u_\sigma\|_\infty + \max_{\omega\in I}|g_\omega| \le C(\|u_\sigma\|_{L^2(\Omega)} + 1) \le C, \tag{2.47}$$

we get

$$\begin{aligned}(S_h^*\lambda_{h,u_\sigma} - S^*\lambda_{h,u_\sigma}, u_\sigma - u_{\sigma,h}) &\le C\|S_h^*\lambda_{h,u_\sigma} - S^*\lambda_{h,u_\sigma}\|_{L^2(\Omega)}\|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &\le Ch^{2-\frac{n}{2}}\|\lambda_{h,u_\sigma}\|_{\mathcal{M}(\Omega)}\|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &\le Ch^{2-\frac{n}{2}}\|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \end{aligned} \tag{2.48}$$

with C independent of σ and h. By (2.26) we have

$$\begin{aligned}(S^*\lambda_{h,u_\sigma} - S^*\lambda_{u_\sigma}, u_\sigma - u_{\sigma,h}) &\le C\|S^*\lambda_{h,u_\sigma} - S^*\lambda_{u_\sigma}\|_{L^2(\Omega)}\|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &\le C\|\lambda_{h,u_\sigma} - \lambda_{u_\sigma}\|_{\mathcal{M}(\Omega)}\|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &= C\left(\sum_{\omega\in I}|S_h u_\sigma(\omega) - Su_\sigma|\right)\|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &\le C\|S_h u_\sigma - Su_\sigma\|_{L^\infty(\Omega)}\|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &\le Ch^{2-\frac{n}{2}}\|u_\sigma\|_{L^2(\Omega)}\|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \\ &\le Ch^{2-\frac{n}{2}}\|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)} \end{aligned} \tag{2.49}$$

with C independent of σ and h. So

$$(S^*\lambda_{h,u_\sigma} - S^*\lambda_{u_\sigma}, u_\sigma - u_{\sigma,h}) \le Ch^{2-\frac{n}{2}}\|u_\sigma - u_{\sigma,h}\|_{L^2(\Omega)},$$

and using this in (2.46) completes the proof. □

Combining Lemmas 2.21 and 2.22 gives Theorem 2.16. A consequence of the theorem is that in 2 dimensions by taking $\sigma = h^2$ we can get an $O(h)$ error estimate for the explicitly discretised problem ($\mathrm{M1}_h$). In this case the state is on a triangulation of size $O(h)$ and the control is on a triangulation of size $O(h^2)$ (i.e. the control space on a finer triangulation than the state). Even though a finer triangulation is involved, the PDEs are posed on the state space to it is reasonable to think of this error estimate as $O(h)$.

Note that Theorem 2.17 can be proved using the same sequence of calculations and bounds as Lemma 2.22. To see this observe that if we replace $U_{ad,\sigma}$ by U_{ad} in both (2.37) and ($\mathrm{M1}_h$), then u_σ solves the continuous problem (2.11) and $u_{\sigma,h}$ solves the implicitly discretised problem ($\mathrm{M2}_h$).

Remark 2.23. *As we noted in Remark 2.15, sometimes* ($\mathrm{M1}_h$) *is equivalent to* ($\mathrm{M2}_h$). *In these cases (e.g. when there are no active control constraints and* $V_h \subset U_\sigma$*) Theorem 2.17 gives error estimates for* ($\mathrm{M1}_h$). *This observation proves Corollary 2.18. In particular it gives an estimate for* ($\mathrm{M1}_h$) *when* $n = 3$ *without control constraints, which Theorem 2.16 does not provide.*

Remark 2.24. *Using this approach to the numerical analysis, the error estimate analogous to Theorem 2.16 for a control problem with an* $L^2(\Omega)$ *fidelity term (instead of one containing point evaluations) is* $O(\sigma + h^2)$ *(see [Casas and Tröltzsch, 2003]).*

2.4.2 Approach 2

This error analysis is based on the technique used in [Deckelnick and Hinze, 2007] and [Leykekhman et al., 2013]. The approach applies to the implicit discretisation ($\mathrm{M2}_h$) (see (2.34)), and therefore also to the explicit discretisation ($\mathrm{M1}_h$) (see (2.31)) when these discrete problems are equivalent (see Remark 2.15). However it does not apply to ($\mathrm{M1}_h$) in general.

The key ingredient of Approach 2 is bounding the difference between the continuous reduced objective functional applied to the discrete and continuous optimal controls, and similarly for the discrete reduced objective functional. Instead of needing error estimates for the control-to-state operator and its adjoint, which were required to prove Theorem 2.17, this approach only uses the strong supremum norm error estimate (2.27). It also does not require the manipulation of measures. As a result this approach is mathematically simpler than Approach 1, but it still allows us to prove the same error estimate as in Theorem 2.17 (modulo ε).

Theorem 2.25. *Let* u *be a solution of (2.11) and* u_h *be a solution of* ($\mathrm{M2}_h$) *(see (2.34)). Then for any* $\varepsilon > 0$*,*

$$\|u - u_h\|_{L^2(\Omega)} \leq C(\varepsilon) h^{2-\frac{n}{2}-\varepsilon}$$

with C *independent of* h*.*

Proof. First observe that

$$\begin{aligned}\hat{J}(u_h) - \hat{J}(u) =& \frac{1}{2}\sum_{\omega\in I}(Su_h - Su)(\omega)^2 + \frac{\nu}{2}\|u_h - u\|^2_{L^2(\Omega)} \\ &+ \sum_{\omega\in I}(Su_h - Su)(Su - g_\omega)(\omega) + \nu(u, u_h - u) \\ \geq& \frac{1}{2}\sum_{\omega\in I}(Su_h - Su)(\omega)^2 + \frac{\nu}{2}\|u_h - u\|^2_{L^2(\Omega)}, \end{aligned} \tag{2.50}$$

since the optimality conditions imply that

$$\sum_{\omega\in I}(Su_h - Su)(Su - g_\omega)(\omega) = a(Su_h - Su, p) = (u_h - u, p) \geq -\nu(u_h - u, u).$$

Similarly

$$\hat{J}_h(u) - \hat{J}_h(u_h) \geq \frac{1}{2}\sum_{\omega\in I}(S_h u_h - S_h u)(\omega)^2 + \frac{\nu}{2}\|u_h - u\|^2_{L^2(\Omega)}. \tag{2.51}$$

Note that the final inequality in this calculation holds for ($\mathrm{M2}_h$) but not for ($\mathrm{M1}_h$) without additional assumptions.

So combining (2.50) and (2.51) we get

$$\begin{aligned}\nu\|u - u_h\|^2_{L^2(\Omega)} &\leq \hat{J}(u_h) - \hat{J}(u) + \hat{J}_h(u) - \hat{J}_h(u_h) \\ &\leq \left|\hat{J}(u) - \hat{J}_h(u)\right| + \left|\hat{J}(u_h) - \hat{J}_h(u_h)\right|. \end{aligned} \tag{2.52}$$

We can bound each of the terms on the right hand side of this inequality. Note that

$$\begin{aligned}\left|\hat{J}(u) - \hat{J}_h(u)\right| &= \left|\frac{1}{2}\sum_{\omega\in I}(Su(\omega) - g_\omega)^2 - \frac{1}{2}\sum_{\omega\in I}(S_h u(\omega) - g_\omega)^2\right| \\ &= \left|\frac{1}{2}\sum_{\omega\in I}(Su - S_h u)(Su - g_\omega + S_h u - g_\omega)(\omega)\right| \\ &\leq C\|Su - S_h u\|_\infty(\|Su\|_\infty + \|S_h u\|_\infty + \max_{\omega\in I}|g_\omega|) \\ &\leq C\|Su - S_h u\|_\infty(\|u\|_{L^2(\Omega)} + 1). \end{aligned}$$

So (2.27) gives that for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$,

$$\begin{aligned}\left|\hat{J}(u) - \hat{J}_h(u)\right| &\leq C(q')h^{3-\frac{n}{q'}}\|u\|_{W_0^{1,q'}(\Omega)}(\|u\|_{L^2(\Omega)} + 1) \\ &\leq C(q')h^{3-\frac{n}{q'}}. \end{aligned} \tag{2.53}$$

In the same way we get that for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$,

$$\begin{aligned}\left|\hat{J}(u_h) - \hat{J}_h(u_h)\right| &\leq C\|Su_h - S_h u_h\|_\infty(\|Su_h\|_\infty + \|S_h u_h\|_\infty + \max_{\omega\in I}|g_\omega|) \\ &\leq C(q')h^{3-\frac{n}{q'}}\|u_h\|_{W_0^{1,q'}(\Omega)}(\|u_h\|_{L^2(\Omega)} + 1). \end{aligned} \tag{2.54}$$

Since $u_h = \mathbb{P}_{[a,b]}(-\frac{1}{\nu}p_h)$ we have $\|u_h\|_{W_0^{1,q'}(\Omega)} \leq C\|p_h\|_{W_0^{1,q'}(\Omega)}$, and the same calculation as in the beginning of Lemma 2.20 gives that

$$\|p_h\|_{W_0^{1,q'}(\Omega)} \leq C(q')$$

independently of h. Combining this, (2.52), (2.53) and (2.54) gives

$$\|u - u_h\|^2_{L^2(\Omega)} \leq C(q')h^{3-\frac{n}{q'}}.$$

Now for any $\varepsilon > 0$ we can set

$$q' = \frac{n}{n - 1 + 2\varepsilon},$$

which completes the proof of the theorem. □

Remark 2.26. *In this proof we used the strong supremum norm estimate (2.27) rather than (2.26). This cannot be used to improve the estimates from Approach 1 in Section 2.4.1; supremum norm estimates are not used in Lemma 2.21, and in Lemma 2.22 we can improve the bound in (2.49) but the error would still be dominated by the $h^{2-\frac{n}{2}}$ term in (2.48).*

2.4.3 Forcing term

We did not include a forcing term in our write up in order to simplify the presentation. However all the results we have proved still hold if we include a forcing term f in the state equation with the regularity $f \in W_0^{1,q'}(\Omega)$ for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$. In particular, if we replace (2.3) by

$$a(y, v) = (\eta + f, v) \quad \forall v \in H_0^1(\Omega), \tag{2.55}$$

and consider a control problem of the form

$$\begin{aligned} \min \quad & J(y,\eta) := \frac{1}{2}\sum_{\omega\in I}(y(\omega)-g_\omega)^2 + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)} \\ \text{over} \quad & C_0(\Omega)\times L^2(\Omega) \\ \text{s.t.} \quad & (2.55) \text{ holds} \\ \text{and} \quad & \eta \in U_{ad} := \{\eta \in L^2(\Omega) : a \le \eta \le b \text{ a.e. in } \Omega\} \end{aligned}$$

with all other assumptions the same as in (2.11). This problem has the reduced form

$$\begin{aligned} \min \quad & \hat{J}(\eta) := \frac{1}{2}\sum_{\omega\in I}(S(\eta+f)(\omega)-g_\omega)^2 + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)} \\ \text{over} \quad & \eta \in U_{ad}, \end{aligned} \tag{2.56}$$

where S is as defined previously. For this problem we can construct non-trivial examples with explicitly known solutions (see Section 2.5.2), which we cannot do for the problem without a forcing term. So after extending our theory to include a forcing term we are able to perform some numerical experiments to verify that our error estimates are observed in practice.

The forcing term means that the mapping from η to y defined by the state equation is no longer linear but instead affine. This difference can be handled with only minor modifications to our problem formulations and proofs, which we now mention: The optimal control problem with forcing still has a unique solution (see e.g. Theorem 1.45 in [Hinze et al., 2009]). Corollary 1.3 in [Hinze et al., 2009] gives that u solves (2.56) if and only if u solves (2.13) with Su replaced by $S(u+f)$ i.e. for all $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$ there exist $p \in W_0^{1,q'}(\Omega)$ such that

$$\begin{aligned} & u \in U_{ad}, \quad (p-\nu u, v-u) \ge 0 && \forall v \in U_{ad}, \\ & a(v,p) = \sum_{\omega\in I}(S(u+f)(\omega)-g_\omega)v(\omega) && \forall v \in W_0^{1,q}(\Omega). \end{aligned}$$

The same reasoning applies to the discrete problems and their optimality conditions with the obvious modifications. In particular the optimality conditions for the discrete problem ($\mathrm{M2}_h$) (see (2.34)) with the inclusion of the forcing term are: There exists a $p_h \in V_h$ such that

$$u_h \in U_{ad}, \quad (p_h + \nu u_h, v - u_h) \ge 0 \qquad \forall v \in U_{ad}, \tag{2.57a}$$

$$a(v_h, p_h) = \sum_{\omega\in I}(S_h(u_h+f)(\omega)-g_\omega)v_h(\omega) \qquad \forall v_h \in V_h. \tag{2.57b}$$

Theorems 2.16, 2.25 and 2.17 still hold with same methods of proof; the f term slightly alters the calculations but does not cause problems, since it follows immediately from the supremum norm error estimate (2.27) that for $\eta \in W_0^{1,q'}(\Omega)$ with $q' \in (\frac{2n}{n+2}, \frac{n}{n-1})$,

$$\|(S - S_h)(\eta + f)\|_\infty \leq C(q')h^{3-\frac{n}{q'}}\|\eta + f\|_{W_0^{1,q'}(\Omega)}.$$

2.5 Numerical results

In this section we develop a numerical method for solving ($\mathrm{M2}_h$) with modification to include a forcing term (see (2.57)) and show that the a priori $L^2(\Omega)$ error estimates that we proved for this discrete problem are numerically realised. In order to do this we solve simple examples of the optimal control problems with explicitly known solutions. We also include some simulations for more interesting problems for which the exact solutions are not known.

2.5.1 Numerical method

We only develop a numerical method for solving ($\mathrm{M2}_h$) because we are able to prove better error estimates for this discrete problem. In particular, we do not have an error estimate for ($\mathrm{M1}_h$) when $n = 3$ with control constraints. Perhaps such an estimate could be proved in other ways, but we cannot easily experimentally investigate if it holds either; we only have explicit solutions (which allow us to reliably test error estimates) when there are no active control constraints. We will now describe the numerical method.

If u_h solves (2.57), then by substituting $u_h = \mathbb{P}_{[a,b]}(-\frac{1}{\nu}p_h)$ we get that the state $y_h := S_h u_h \in V_h$ and the adjoint variable $p_h \in V_h$ solve

$$\begin{pmatrix} a(y_h, v_h) - (-\frac{1}{\nu}p_h + (a + \frac{1}{\nu}p_h)^+ - (-\frac{1}{\nu}p_h - b)^+ - f, v_h) \\ a(w_h, p_h) - \sum_{\omega \in I}(y_h(\omega) - g_\omega)w_h(\omega) \end{pmatrix} = 0 \qquad (2.58)$$

for all $v_h, w_h \in V_h$. Here v^+ denotes the nonnegative part of v i.e. $\max(0, v)$. Once this problem has been solved, the u_h solving (2.35a) can easily be determined from p_h by setting $u_h = \mathbb{P}_{[a,b]}\big(-\frac{1}{\nu}p_h\big)$. We will now describe a numerical method for solving (2.58) with and without control constraints.

No control constraints

In the case of no control constraints ($b = -a = \infty$) the nonlinear $\max(0, \cdot)$ terms drop out, leaving a linear problem. Let $y_h = \sum_{z \in \mathcal{N}} y_z \varphi_z$ and $p_h = \sum_{z \in \mathcal{N}} p_z \varphi_z$, where φ_z are the usual nodal basis functions of V_h (defined by $\varphi_z(\bar{z}) = \delta_{z\bar{z}}$ for $\bar{z} \in \mathcal{N}$, where $\delta_{z\bar{z}}$ denotes the Kronecker delta and $\mathcal{N}$ is the set of interior vertices of the triangulation), and y_z and p_z are the coefficients corresponding to the basis functions. As we have no control constraints, testing (2.58) with $v_h = \varphi_z$ and $w_h = \varphi_{\bar{z}}$ for all $z, \bar{z} \in \mathcal{N}$ leads to a system of linear equations of real variables. In particular, let $\bar{y}$ and $\bar{p}$ be vectors of coefficients defined by $\bar{y}_z = y_z$ and $\bar{p}_z = p_z$ for $z \in \mathcal{N}$ i.e. use the set of interior vertices as an index. Then we can solve (2.58) by solving the system of linear equations

$$\begin{pmatrix} A & \frac{1}{\nu} M \\ -\sum_{\omega \in I} M_\omega & A \end{pmatrix} \begin{pmatrix} \bar{y} \\ \bar{p} \end{pmatrix} = \begin{pmatrix} \bar{F} \\ -\sum_{\omega \in I} \bar{G}_\omega \end{pmatrix},$$

where matrices A, M and M_ω and vectors $\bar{F}$ and $\bar{G}_\omega$ are defined by

$$\begin{aligned} &A_{z\bar{z}} = a(\varphi_z, \varphi_{\bar{z}}), \quad M_{z\bar{z}} = (\varphi_z, \varphi_{\bar{z}}), \quad (M_\omega)_{z\bar{z}} = \varphi_z(\omega)\varphi_{\bar{z}}(\omega) && \forall z, \bar{z} \in \mathcal{N}, \\ &\bar{F}_z = (f, \varphi_z), \quad (\bar{G}_\omega)_z = g_\omega \varphi_z(\omega) && \forall z \in \mathcal{N}. \end{aligned}$$

As the basis functions φ_z are piecewise linear with small support, the integrals that form the elements of the matrices and vectors are straightforward to compute, assuming A and f have a simple form (or else numerical integration of some terms may be required, which we discuss later). The matrix in this system of equations is sparse and so the system can be solved efficiently.

Control constraints

In the case of control constraints the nonlinear $\max(0, \cdot)$ terms mean that we can no longer use the above approach to construct a linear system of equations of real variables. Instead we will solve the problem iteratively using a Newton-type method. Let $F_h : V_h \times V_h \to V_h^* \times V_h^*$ with $F_h(y_h, p_h)(w_h, v_h)$ defined by the left hand side of (2.58). Then it can be written as

$$F_h(y_h, p_h) = 0 \quad \text{in } V_h^* \times V_h^*. \tag{2.59}$$

The $\max(0, \cdot)$ terms mean that F_h is not Fréchet differentiable. However we can apply a generalised Newton method called the semismooth Newton method (see e.g. [Ulbrich, 2002] and [Hintermüller and Kopacka, 2009]). This amounts to applying

the Newton method in the usual way but taking the derivative of $\max(0,x)$ to be

$$\max'(0,x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0. \end{cases}$$

So we take an initial guess y_h^0, p_h^0 then apply until the convergence the semismooth Newton iteration

$$\begin{pmatrix} y_h^{n+1} \\ p_h^{n+1} \end{pmatrix} = \begin{pmatrix} y_h^n \\ y_h^n \end{pmatrix} + \begin{pmatrix} \delta y_h \\ \delta p_h \end{pmatrix},$$

where δy_h, δp_h solve

$$\begin{aligned} &F_h'(y_h^n, p_h^n)(\delta y_h, \delta p_h) \\ &= \begin{pmatrix} a(\delta y_h, \cdot) - \frac{1}{\nu}\Big(\big(-1 + \max'(0, a + \frac{1}{\nu} p_h^n) + \max'(0, -\frac{1}{\nu} p_h^n - b)\big)\delta p_h, \cdot \Big) \\ a(\cdot, \delta p_h) - \sum_{\omega \in I} \delta y_h(\omega)(\cdot)(\omega) \end{pmatrix} \\ &= -F_h(y_h^n, p_h^n). \end{aligned} \tag{2.60}$$

Note that if we have no control constraints the first Newton iteration is equivalent to solving (2.58).

As before we can represent δy_h and δp_h as a sum of basis functions weighted by coefficients, and testing (2.60) with the basis functions allows us to construct the following system of linear equations of real variables:

$$\begin{pmatrix} A & \frac{1}{\nu} M_c \\ -\sum_{\omega \in I} M_\omega & A \end{pmatrix} \begin{pmatrix} \delta \bar{y} \\ \delta \bar{p} \end{pmatrix} = \begin{pmatrix} \bar{R}_1 \\ \bar{R}_2 \end{pmatrix},$$

where

$$\begin{aligned} &(M_c)_{z\bar{z}} := (c(x)\varphi_z, \varphi_{\bar{z}}) && \forall z, \bar{z} \in \mathcal{N}, \\ &\qquad \text{with } c(x) := 1 - \max'(0, a + \frac{1}{\nu} p_h^n(x)) - \max'(0, -\frac{1}{\nu} p_h^n(x) - b), \\ &(\bar{R}_1)_z := -(F_h(y_h^n, p_h^n)_1, \varphi_z), \quad (\bar{R}_2)_z := -(F_h(y_h^n, p_h^n)_2, \varphi_z) && \forall z \in \mathcal{N}. \end{aligned}$$

Note that since $p_h^n(x)$ is piecewise linear, the integrals of functions such as $\max'(0, a + \frac{1}{\nu} p_h^n)\varphi_z\varphi_{\bar{z}}$ can be computed exactly. In practice we instead approximate this using a numerical quadrature i.e. instead of $(c(x)\varphi_z, \varphi_{\bar{z}})$ we will compute $Q(c(x)\varphi_z\varphi_{\bar{z}})$, where

$$Q(\eta) := \sum_{T \in T_h} Q_T(\eta), \quad Q_T(\eta) := \sum_{q=1}^{K} \hat{w}_q \, |DF_T(\hat{x}_q)| \, \eta(F_T(\hat{x}_q)).$$

Here $\{(\hat{w}_q, \hat{x}_q)\}_{q=1}^K$ is a collection of K pairs of weights and points on a reference element $\hat{T}$ and F_T is the reference mapping between $\hat{T}$ and T. We will use a Gaussian quadrature of high order (large K), so $Q(\eta) \approx \int_\Omega \eta(x)\mathrm{d}x$. We will also use this quadrature rule to approximate f as it may have a form that makes it complicated to integrate by hand. The moderately large error from our discretisation should dominate the smaller error from Gaussian quadrature (as it has good approximation properties), so we do not expect using quadrature to affect the $L^2(\Omega)$ error we observe in practice. Note that using quadrature means that we are not solving (2.58) but rather a close approximation. Although using quadrature is not strictly necessary, the implementation without would require us to do additional calculations by hand, particularly in 3 dimensions. In comparison, there is built in support for numerical quadrature in many finite element software packages.

Define the product space norm for $(z_1, z_2) \in Z \times Z$, where Z is a normed vector space, by $\|(z_1, z_2)\|_Z = \sqrt{\|z_1\|_Z^2 + \|z_2\|_Z^2}$. For $z \in H^{-1}(\Omega)$ let $w \in H_0^1(\Omega)$ be defined by

$$(\nabla w, \nabla v) = \langle z, v\rangle_{H^{-1}(\Omega)} \quad \forall v \in H_0^1(\Omega).$$

Then

$$\begin{aligned}\|z\|_{H^{-1}(\Omega)} &= \sup_{v\in H_0^1(\Omega)} \frac{\langle z, v\rangle_{H^{-1}(\Omega)}}{\|v\|_{H_0^1(\Omega)}} \\ &= \sup_{v\in H_0^1(\Omega)} \frac{(\nabla w, \nabla v)}{\|v\|_{H_0^1(\Omega)}} \\ &= \|w\|_{H_0^1(\Omega)}.\end{aligned}$$

This motivates us to iterate the Newton method until the stopping criterion $\|F_h(y_h, p_h)\|_Z$ is small, where for $z_h \in V_h^*$ we define $\|z_h\|_Z := \|w_h\|_{H_0^1(\Omega)}$ with

$$w_h \in V_h, \quad (\nabla w_h, \nabla v_h) = \langle z_h, v_h\rangle_{V_h^*} \quad \forall v_h \in V_h.$$

Note that if $\|F_h(y_h, p_h)\|_Z = 0$ then (y_h, p_h) is the solution to (2.59). The algorithm we use is stated precisely in Algorithm 1 below.

Newton type methods typically offer local superlinear convergence. We do not prove this, but we note in Section 2.5.5 that our algorithm is very effective in practice. On all the problems we tested it provided quadratic mesh independent convergence to the solution even with the bad initial iterate of $(0, 0)$.

Algorithm 1 Newton method

Input: T_h, y_h^0, p_h^0 and DATA $= (\Omega, \nu, f, a, b, I, \{y_w\}_{\omega \in I})$ ▷ $(y_h^0, p_h^0) = (0, 0)$

1: **while** $\|F_h(y_h^k, p_h^k)\|_Z > \delta$ **do** ▷ $\delta = 1e-8$
2: Compute $(\delta y_h, \delta p_h)$ by solving (2.60): $F_h'(y_h^k, p_h^k)(\delta y_h, \delta p_h) = -F_h(y_h^k, p_h^k)$.
3: $(y_h^{k+1}, p_h^{k+1}) \leftarrow (y_h^k, p_h^k) + (\delta y_h, \delta p_h)$
4: $k \leftarrow k + 1$
5: **end while**
6: **return** y_h^k, p_h^k

Implementation

As we remarked above, in the case of no control constraints the first iteration of the Newton method solves (2.58). So rather than implementing two different numerical methods, we also use Algorithm 1 to solve the problem when there are no control constraints.

We implemented Algorithm 1 in the Distributed and Unified Numerics Environment (DUNE) using DUNE-FEM [Blatt and Bastian, 2007, 2008; Bastian et al., 2008b,a, 2011; Dedner et al., 2010, 2011]. This environment has the advantage that once an algorithm has been implemented, it is straightforward to change features of the implementation that would usually be fixed. For solving the linear systems for each iteration of the Newton method we used the biconjugate gradient stabilised method with an incomplete LU factorisation or Gauss-Seidel preconditioner.

2.5.2 Exact solutions

We can construct an exact solution for a simple example of the optimal control problem in dimensions 2 and 3 without control constraints. This allows us to verify our error estimates. The key fact we will use to do this is that fundamental solutions of the Laplace equation $-\Delta y = \delta_{x'}$ are given by

$$\begin{cases} -\frac{1}{2\pi} \log|x - x'| + C & n = 2, \\ \frac{1}{4\pi|x-x'|} + C & n = 3. \end{cases}$$

So take $\Omega = B_1(0)$, the open unit ball in $\mathbb{R}^n$ centred at the origin, and $I = \{0\}$. Then

$$p(x) = \begin{cases} -\frac{1}{2\pi} \log|x| \, (y(0) - g_0) & n = 2 \\ \frac{1}{4\pi}(\frac{1}{|x|} - 1)(y(0) - g_0) & n = 3 \end{cases}$$

is the unique p solving (2.13b), and $u = -\frac{1}{\nu}p$ (as we have no control constraints). Note that u and p are unbounded, however they are still $L^2(\Omega)$ functions. To see

this note that converting to polar and spherical coordinates we have

$$\int_\Omega (\log|x|)^2 \mathrm{d}x = \int_0^{2\pi}\int_0^1 (\log r)^2 r \,\mathrm{d}r\,\mathrm{d}\theta < \infty,$$
$$\int_\Omega \frac{1}{|x|^2}\mathrm{d}x = \int_0^{2\pi}\int_0^{\pi}\int_0^1 \sin\theta \,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\varphi < \infty.$$

We can now set y to be any function satisfying the boundary conditions (e.g. $y(x) = \cos(\frac{\pi|x|}{2})$), and take $f = -\Delta y - u$. We also set $\nu = 1$ and $g_0 = y(0) - 1$ to simplify the problem and exact solution further.

2.5.3 2D numerical results

Motivated by the above construction take $\Omega = B_1(0)$, $A = -\Delta$, $I = \{0\}$, $g_0 = 0$, $b = -a = \infty$, $\nu = 1$, and

$$f = \frac{\pi}{4}\left(\frac{2}{|x|}\sin\left(\frac{\pi\,|x|}{2}\right) + \pi\cos\left(\frac{\pi\,|x|}{2}\right)\right) - \frac{1}{2\pi}\log|x|\,.$$

Then the solution to the control problem is

$$u(x) = -p(x) = \frac{1}{2\pi}\log|x|\,,$$
$$y(x) = \cos\left(\frac{\pi\,|x|}{2}\right).$$

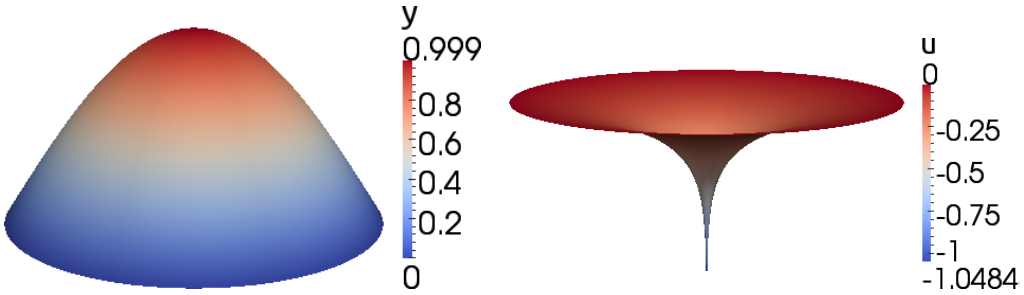
This solution is interesting because the control is singular (infinite) at the prescribed point $(0,0)$ but it is still an $L^2(\Omega)$ function. We solve this problem numerically using the numerical method outlined in Section 2.5.1, giving Figure 2.2. Note that the solution to the discrete problem must be bounded, even though it is approximating am unbounded function. As a result, the magnitude of the spike in u_h notably increases as the triangulation is refined (but $\|u_h\|_{L^2(\Omega)}$ is stable).

The computed $L^2(\Omega)$ errors are in Table 2.2, where $\|u - u_h\|_{L^2(\Omega)}$ is approximated using a Gaussian quadrature rule of high order, and the experimental order of convergence is defined by

$$\mathrm{EOC}_h = \frac{\log(\|u - u_{h/2}\|_{L^2(\Omega)}/\|u - u_h\|_{L^2(\Omega)})}{\log 2}.$$

The data suggest order h convergence for this problem, which agrees with the estimate we proved in Theorem 2.25

The solution of a more interesting problem including control constraints and more evaluation points is shown on the left hand side of Figure 2.3. It appears that



(a) y_h (b) $u_h = -p_h$

Figure 2.2: The radially symmetric solution to our 2D problem with explicitly known solution.

h	#DoFs	$\|u - u_h\|_{L^2(\Omega)}$	EOC_h
0.5	25	0.03258	-
0.25	81	0.0160362	1.0226543
0.125	289	0.00787259	1.0264221
0.0625	1089	0.00389451	1.0153965
0.03125	4225	0.00193778	1.0070370
0.015625	16641	0.000966977	1.0028513
0.0078125	66049	0.00048313	1.0010701

Table 2.2: EOCs for the 2D problem with explicitly known solution (see Figure 2.2).

p_h is approximating an unbounded p, though $\|p_h\|_{L^2(\Omega)}$ is bounded. However u is certainly bounded due to the control constraints. We do not know the exact solution to this problem so we cannot calculate the error exactly. However we can calculate an approximate order of convergence by comparing to the solution on a very fine triangulation i.e. $\tilde{u} = u_{h_{\text{fine}}}$ with $h_{\text{fine}} = 0.00276214$, which corresponds to 263169 DOFs. So we instead compute

$$\mathrm{EOC}_h = \frac{\log(\|\tilde{u} - u_{h/2}\|_{L^2(\Omega)} / \|\tilde{u} - u_h\|_{L^2(\Omega)})}{\log 2} \tag{2.61}$$

for $h \gg h_{\text{fine}}$. We ensure that the fine triangulation is a refinement of the coarser triangulations, so the $L^2(\Omega)$ errors can be computed accurately using a high order Gaussian quadrature. These approximate EOCs can be seen in Table 2.3. They agree with the error estimate we proved for the case of active control constraints in Theorem 2.25. The slight increase in the EOC for the smallest value of h is expected as we are computing the error against a discrete solution and not the true solution.

On the right hand side of Figure 2.3 we have the solution of the 2D problem we just considered but without the control constraints. We observe that this allows the state to get slightly closer to the prescribed values. In order to get closer still

h	#DoFs	$\|u-u_h\|_{L^2(\Omega)}$	EOC_h
0.353553	25	2.8881	-
0.176777	81	1.51039	0.93520339
0.0883883	289	0.80295	0.91153608
0.0441942	1089	0.409627	0.97100093
0.0220971	4225	0.205786	0.99316598
0.0110485	16641	0.100486	1.0341436

Table 2.3: EOCs for the 2D problem on the left hand side of Figure 2.3, which has control constraints.

h	#DoFs	$\|u-u_h\|_{L^2(\Omega)}$	EOC_h
1	27	0.103658	-
0.5	125	0.0719594	0.52657640
0.25	729	0.0474726	0.60008809
0.125	4913	0.0322929	0.55587806
0.0625	35937	0.0225399	0.51873589

Table 2.4: EOCs to our 3D problem with explicitly known solution (see Figure 2.5).

we would need to decrease ν. Figure 2.4 shows a more interesting example with $\nu = 1e-4$ (i.e. very small). As a result the state takes values very close to the prescribed values, and overshoots the value 1 on parts of the domain in order to achieve this.

2.5.4 3D numerical results

Similarly take $\Omega = B_1(0)$, $A = -\Delta$, $I = \{0\}$, $g_0 = 0$, $b = -a = \infty$, $\nu = 1$, and

$$f = \frac{\pi}{4}\left(\frac{4}{|x|}\sin\left(\frac{\pi\,|x|}{2}\right) + \pi\cos\left(\frac{\pi\,|x|}{2}\right)\right) + \frac{1}{4\pi}\Big(\frac{1}{|x|} - 1\Big).$$

Then the solution to the control problem is

$$\begin{aligned} u(x) &= -p(x) = -\frac{1}{4\pi}\Big(\frac{1}{|x|} - 1\Big), \\ y(x) &= \cos\left(\frac{\pi\,|x|}{2}\right). \end{aligned}$$

This solution can be seen in Figure 2.5. We observe order $\sqrt{h}$ convergence (see Table 2.4), which again agrees with the estimate we proved in Theorem 2.25.

h	# iterations
0.0883883	3
0.0441942	3
0.0220971	3
0.0110485	3
0.00552427	3

Table 2.5: Number of Newton iterations needed for a given h for the problem with control constraints in Figure 2.3 (i.e. a nonlinear problem) with the initial iterate $(0, 0)$.

Iteration k	$\|F_h(u_h^k)\|_Z$	EOC_k
0	0.00285272	0
1	4.38339×10^{-5}	0.85936125
2	1.21172×10^{-6}	2.4577166
3	1.79175×10^{-10}	0

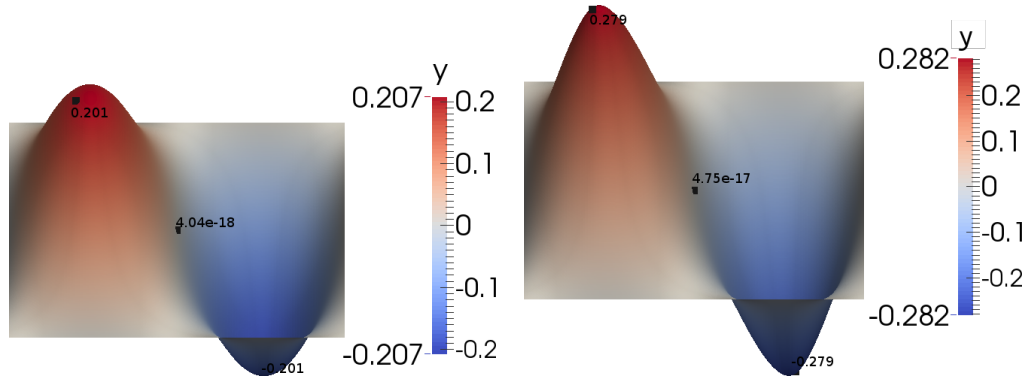
Table 2.6: Convergence rate of Newton method.

2.5.5 Mesh independence

We finish this chapter by justifying the effectiveness of our numerical method. When we have no control constraints the problem is linear and the Newton method always finds the exact solution in a single iteration. When we have control constraints the problem is nonlinear and we still have good mesh independence properties; the number of Newton iterations needed for convergence does not increase as h is decreased (see Table 2.5).

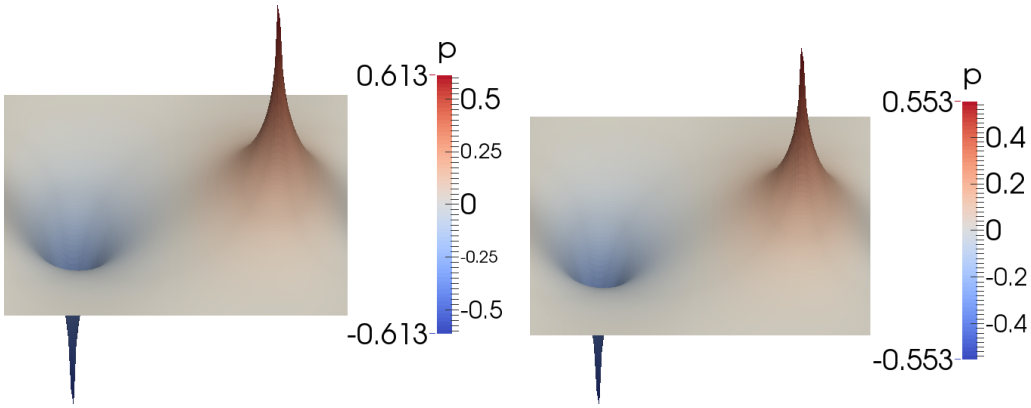
We also observe quadratic convergence of the Newton method on average. See Table 2.6 for the residuals of the Newton method for the control constrained example from Figure 2.3. In the table

$$\mathrm{EOC}_k := \frac{\log(\delta_{k+1}/\delta_k)}{\log(\delta_k/\delta_{k-1})}, \quad \delta_k := \|F_h(y_h^k, p_h^k)\|_{H^{-1}(\Omega)}. \tag{2.62}$$



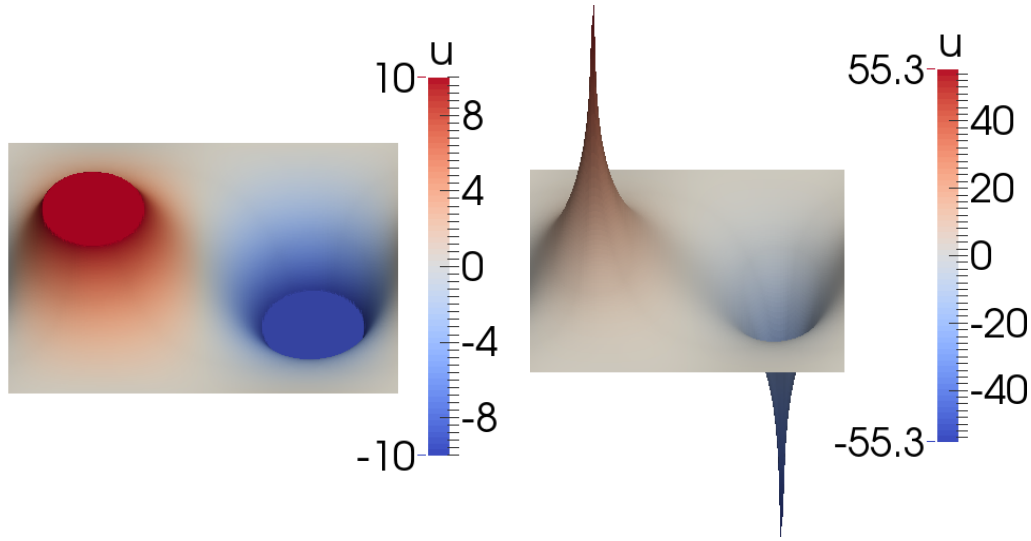
(a) y_h with $b = -a = 10$.

(b) y_h with $b = -a = \infty$.



(c) p_h with $b = -a = 10$.

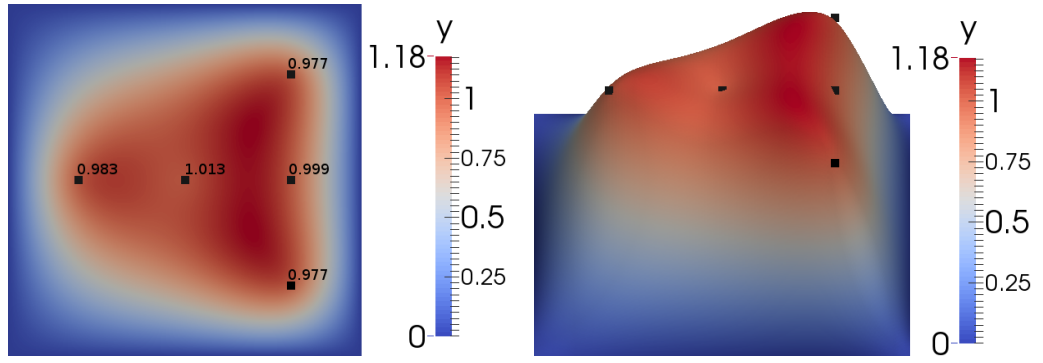
(d) p_h with $b = -a = \infty$.



(e) $u_h = \mathbb{P}_{[a,b]}(-\frac{1}{\nu}p_h)$ with $b = -a = 10$.

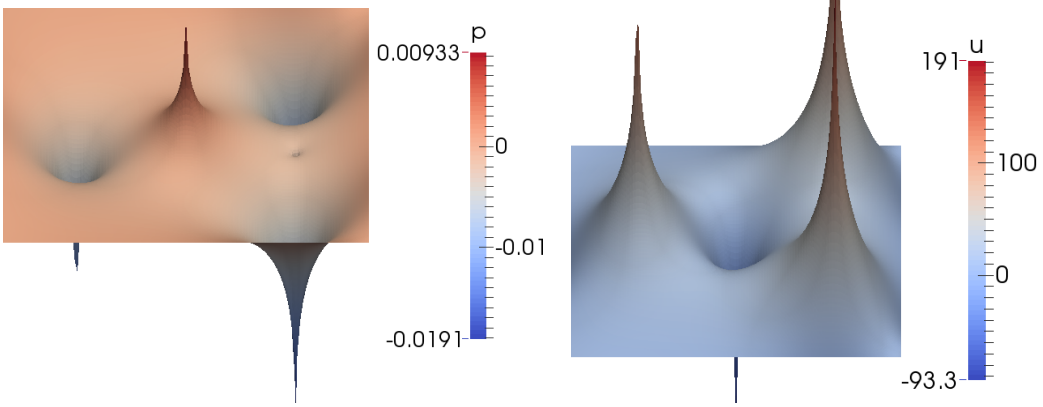
(f) $u_h = \mathbb{P}_{[a,b]}(-\frac{1}{\nu}p_h)$ with $b = -a = \infty$.

Figure 2.3: Solutions for $\Omega = (0,1)^2$, $A = -\Delta$, $f = 0$, $I = \{(0.2, 0.5), (0.5, 0.5), (0.8, 0.5)\}$, $y_{(0.2,0.5)} = 1$, $y_{(0.5,0.5)} = 0$, $y_{(0.8,0.5)} = -1$, and $\nu = 1e-2$. The solution on the right has $b = -a = 10$ and the solution on the left has no control constraints ($b = -a = \infty$). The scale on figures that are side by side is the same. The black dots mark the locations of the points in I and the numbers give the value of y_h at these points.



(a) y_h

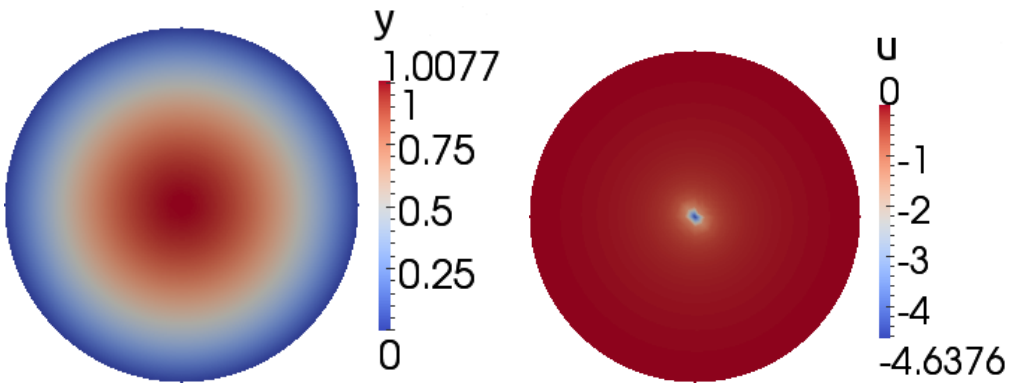
(b) y_h



(c) p_h

(d) $u_h = -\frac{1}{\nu} p_h$

Figure 2.4: Solution to a more interesting example with $\Omega = (0,1)^2$, $A = -\Delta$, $f = 0$, $I = \{(0.2, 0.5), (0.5, 0.5), (0.8, 0.2), (0.8, 0.5), (0.8, 0.8)\}$, $g_\omega = 1$ for all $\omega \in I$, $\nu = 1e-4$, and $b = -a = \infty$.



(a) y_h (b) $u_h = -p_h$

Figure 2.5: A slice passing through the origin of the radially symmetric solution to our 3D problem with explicitly known solution.

Chapter 3

Optimal control of elliptic PDEs on surfaces of codimension 1

In this chapter we continue our study of optimal control of elliptic PDEs on sets of measure zero by considering control on surfaces of codimension 1, also known as hypersurfaces. In particular, we control the state to be close to prescribed values along a curve in 2D or a surface in 3D. So for a bounded domain $\Omega \subset \mathbb{R}^n$ ($n = 2$ or 3) and an $n-1$ dimensional surface $\Gamma \subset \Omega$ we consider the problem:

$$\min \frac{1}{2}\int_\Gamma (y - g_\Gamma)^2 \mathrm{d}A + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)} \tag{3.1}$$

subject to the state equation

$$\begin{aligned} Ay &= \eta \quad \text{in } \Omega \\ y &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{3.2}$$

and the control constraints

$$a \leq \eta \leq b.$$

Here $g_\Gamma : \Gamma \to \mathbb{R}$ is the desired state on Γ, and the other notation is the same as in Chapter 2. We will formulate this problem precisely in Section 3.2.

The motivation for the surface fidelity term is that in some applications we may only care about the state being close to given values on a small part of the domain. Controlling the state using a distributed norm over the whole domain gives weaker control on the surface. Instead of this fidelity term we could use state constraints to force the state to take certain values, however this would lead to an optimal control with very high cost. The surface fidelity term allows for a

compromise between how close the state is to the desired values on the surface and the cost of the control.

We have not seen the surface fidelity term previously used in the optimal control context in the literature, however other problems have been considered where the state is controlled on small sets. In Chapter 2 we do a detailed numerical analysis of finite element discretisations of an elliptic optimal control problem with a point fidelity term. This chapter cites other related literature. [Brett et al., 2013] and Chapter 4 develop an adaptive finite element method for a point control problem with a variational inequality state constraint.

In comparison to point control problems the difficulty of our problem is not the low regularity of the adjoint variable; it belongs to $H_0^1(\Omega)$ and standard literature (e.g. [Tröltzsch, 2010]) provides the necessary background for the analysis. The difficulty is that in order to pose a discrete problem that can be solved computationally, we may need to formulate the discrete problem with an approximation of Γ, such as a polygonal (for $n = 2$) or polyhedral (for $n = 3$) approximation. This complicates the numerical analysis, and estimating the error caused by approximating Γ forces us to introduce theory usually associated with finite element methods for PDEs on hypersurfaces, such as that reviewed in [Dziuk and Elliott, 2013]. Note that we do not consider the case of a curve in 3D as this would require additional regularity of the state. In particular, we would need the state to be continuous so the problem is more closely related to that in Chapter 2.

Other related optimal control problems have been considered in the literature. The recent paper [Gong et al., 2014] considers elliptic optimal control problems with controls on lower dimensional manifolds. Their state equation has a similar form to our adjoint equation, and our state equation has a similar form to their adjoint equation. In their discrete problems they also approximate surfaces with polyhedral surfaces. Note that our assumptions on these approximating hypersurfaces are more flexible. In papers such as [Casas et al., 2012] and [Pieper and Vexler, 2013] a problem is considered where the control space is a space of measures. Supremum norm error estimates that are needed when working with state constrained elliptic optimal control problems are useful to us. [Leykekhman et al., 2013] proves error estimates for problems with state constraints at a finite number of points. [Deckelnick and Hinze, 2007] proves error estimates for the case of global (as opposed to point) state constraints, but for a state equation with Neumann boundary conditions. A review of the analysis for standard optimal control problems can be found in [Tröltzsch, 2010] and a review of the numerical analysis can be found in [Hinze et al., 2009].

	$n=2$	$n=3$
Theory	$O(h^{1-\varepsilon})$	$O(h^{\frac{3}{4}})$
Numerics	$O(h)$	-

Table 3.1: The main a priori error estimates proved for $\|u-u_h\|_{L^2(\Omega)}$.

We will define an appropriate finite element discretisation of our problem and prove a priori error estimates for the $L^2(\Omega)$ error in the control. Our discretisation is based on the variational discretisation idea from [Hinze, 2005], as this typically allows for better error estimates. We will prove these error estimates using an approach inspired by [Deckelnick and Hinze, 2007], since we found it to be relatively simple. This will allow us to focus on the new difficulties caused by approximating Γ. We will show numerical results for $n=2$ that agree with our analytical results. We do not solve any examples for $n=3$ as the implementation would be more complicated. Table 3.1 summarises our results, where $\varepsilon>0$ is arbitrary.

In the next section we introduce some notation. In Section 3.2 we formulate the optimal control problem precisely and prove some analytical results. In Section 3.3 we introduce the theory for approximating hypersurfaces and discretise using a finite element method. In Section 3.4 we prove a priori error estimates for the $L^2(\Omega)$ error in the control. In Section 3.5 we show numerical results.

3.1 Notation

Let the domain $\Omega\subset\mathbb{R}^n$ ($n=2$ or 3) be an open bounded domain. Let A be a differential operator satisfying the same assumptions as in Section 2.1, and let $a: H_0^1(\Omega)\times H_0^1(\Omega)\to\mathbb{R}$ be the associated bilinear form. Then as before, given $\eta\in L^2(\Omega)$ there is a unique weak solution $y\in H_0^1(\Omega)$ to (3.2) defined by

$$a(y,v)=(\eta,v)_{L^2(\Omega)}\quad\forall v\in H_0^1(\Omega). \tag{3.3}$$

We will see shortly that in contrast to Chapter 2, it is not necessary for the state to be continuous in order for the objective functional to be well defined. The state belonging to $H_0^1(\Omega)$ is sufficient. This is why we do not make stronger assumptions on the regularity of the domain (though we will in Section 3.3 as it is necessary for the numerical analysis). We define the control-to-state operator $S: L^2(\Omega)\to H_0^1(\Omega)$ by $S\eta:=y$, where y solves (3.3). This operator is linear, and

continuous since testing (3.3) with $v = y$ allows us to deduce

$$\|S\eta\|_{H_0^1(\Omega)} \leq C\|\eta\|_{L^2(\Omega)}.$$

This means we can define the adjoint operator $S^* : H^{-1}(\Omega) \to L^2(\Omega)$ in the usual way by

$$(S^*z, \eta)_{L^2(\Omega)} = \langle z, S\eta\rangle_{H^{-1}(\Omega)} \quad \forall z \in H^{-1}(\Omega), \eta \in L^2(\Omega), \tag{3.4}$$

where $\langle\cdot,\cdot\rangle_{H^{-1}(\Omega)}$ abbreviates the usual duality pairing $\langle z, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = z(v)$. The adjoint operator has the following property.

Lemma 3.1. *For $f \in H^{-1}(\Omega)$, $p = S^*f$ if and only if p satisfies*

$$p \in H_0^1(\Omega), \quad a(v, p) = \langle f, v\rangle_{H^{-1}(\Omega)} \quad \forall v \in H_0^1(\Omega). \tag{3.5}$$

Proof. Suppose (3.5) is true. Then for any $f \in H^{-1}(\Omega)$ and $\eta \in L^2(\Omega)$, testing with $S\eta \in H_0^1(\Omega)$ we get

$$a(S\eta, p) = \langle f, S\eta\rangle_{H^{-1}(\Omega)}.$$

Since $p \in H_0^1(\Omega)$, by the definition of S we have

$$a(S\eta, p) = (\eta, p)_{L^2(\Omega)} = (p, \eta)_{L^2(\Omega)}.$$

Combining these two equalities and recalling that f is arbitrary we get

$$\langle f, S\eta\rangle_{H^{-1}(\Omega)} = (p, \eta)_{L^2(\Omega)} \quad \forall f \in H^{-1}(\Omega), \eta \in L^2(\Omega).$$

Comparing this to the definition of the adjoint we see that $p = S^*f$. Since the adjoint operator is unique, the reverse statement must also hold. This completes the proof. □

Let Γ be a C^2-hypersurface (see e.g. Section 2.2 in [Dziuk and Elliott, 2013] for this definition) for which $\Gamma \subset \Omega$ and there exists an open set $U \subset \Omega$ with a Lipschitz boundary such that $\Gamma \subset \partial U$. Note that we allow Γ to be an open hypersurface (i.e. one that has a boundary) and it may have multiple connected components.

We now give an example of an admissible Γ and a corresponding U.

Example 3.2. *Suppose $n = 2$ and let $\Omega := (0, 1)^2 \subset \mathbb{R}^n$. Consider*

$$\Gamma := \{(0.25 + 0.5t, 0.5) : t \in (0, 1)\}.$$

i.e. Γ *is a straight line (an open hypersurface). Note that* Γ *is orientable, so there is a continuous vector field* $\mu : \Gamma \to \mathbb{R}^n$ *such that* $\mu(c)$ *is a unit normal to* Γ *for all* $c \in \Gamma$. *Therefore we can take*

$$U := \{c + D\mu(c) \in \mathbb{R}^n : c \in \Gamma, 0 < D < \theta\} \subset \Omega$$

with $\theta = 0.1$. *Observe that* $\Gamma \subset \partial U$ *and* U *is an open set with a Lipschitz boundary.*

This construction of U *(for sufficiently small* $\theta > 0$*) works for many choices of hypersurface, including closed hypersurfaces such as*

$$\Gamma := \{(0.5 + 0.25\cos(2\pi t), 0.5 + 0.25\sin(2\pi t) : t \in (0,1)\} \subset (0,1)^2$$

i.e. a circle.

We can define some function spaces on Γ. Denote by $C(\Gamma)$ the set of functions which are continuous on Γ. Let $L^s(\Gamma)$ with $s \in [1, \infty]$ denote the space of functions $v : \Gamma \to \mathbb{R}$ which are measurable with respect to the surface measure $\mathrm{d}A$ (the $n-1$ dimensional Hausdorff measure) and have a finite norm

$$\begin{aligned} \|v\|_{L^s(\Gamma)} &:= \left(\int_\Gamma |v|^s \,\mathrm{d}A\right)^{\frac{1}{s}} && s \in [1, \infty), \\ \|v\|_{L^\infty(\Omega)} &:= \operatorname{ess\,sup} |v| && p = \infty. \end{aligned}$$

These spaces are Banach spaces, and $L^2(\Gamma)$ is a Hilbert space with inner product

$$(v, w)_{L^2(\Gamma)} := \int_\Gamma vw \mathrm{d}A.$$

Since Γ is a C^2 hypersurface we can also define weak derivatives of functions in $L^1(\Gamma)$ and hence Sobolev spaces $H^{k,p}(\Gamma)$. As usual, $H^k(\Gamma)$ is used to denote $H^{k,2}(\Gamma)$. We do not use these Sobolev spaces directly so we leave the reader to refer to e.g. [Dziuk and Elliott, 2013] for the details.

We need to check that we can make sense of $y|_\Gamma$ for $y \in H_0^1(\Omega)$.

Lemma 3.3. *Let* Γ *be a hypersurface satisfying the above assumptions. Then there exists a continuous linear operator* $T : H_0^1(\Omega) \to L^2(\Gamma)$ *such that*

$$Ty = y|_\Gamma \quad \forall y \in H_0^1(\Omega) \cap C(\Omega). \tag{3.6}$$

In particular,

$$\|Ty\|_{L^2(\Gamma)} \leq C\|y\|_{H_0^1(\Omega)} \quad \forall y \in H_0^1(\Omega), \tag{3.7}$$

with C independent of y.

Proof. Let $\tilde{T} : H^1(U) \to L^2(\partial U)$ denote the trace operator for the open set U which has a Lipschitz boundary i.e. the unique continuous linear operator from $H^1(U)$ to $L^2(\partial U)$ such that $\tilde{T}y = y|_{\partial U}$ for all $y \in C^{0,1}(U)$ (see e.g. Theorem 1.5.1.3 in [Grisvard, 1985]). Then we can define $T : H_0^1(\Omega) \to L^2(\Gamma)$ by

$$Ty := \tilde{T}(y|_U)\big|_\Gamma.$$

The linearity of $\tilde{T}$ implies that T is linear. It is also straightforward to see that (3.6) holds. Finally, using the continuity of the linear operator $\tilde{T}$, we get that for $y \in H_0^1(\Omega)$,

$$\|Ty\|_{L^2(\Gamma)} \leq \|\tilde{T}y\|_{L^2(\partial U)} \leq C\|y\|_{H^1(U)} \leq C\|y\|_{H_0^1(\Omega)}$$

with C independent of y. So (3.7) holds and T is continuous. □

For $y \in H_0^1(\Omega)$ we use $y|_\Gamma$ to denote Ty and (when it will not cause confusion) write quantities such as $\|y\|_{L^2(\Gamma)}$ instead of $\|y|_\Gamma\|_{L^2(\Gamma)}$. So with this notation the objective functional (3.1) is well defined.

3.2 Problem formulation

We now formulate the optimal control problem precisely as

$$\begin{aligned}
&\min\ J(y,\eta) := \frac{1}{2}\|y - g_\Gamma\|^2_{L^2(\Gamma)} + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)} \\
&\text{over } H_0^1(\Omega) \times L^2(\Omega) \\
&\text{ s.t. } y = Su \text{ (i.e. (3.3) holds)} \\
&\text{and } \eta \in U_{ad} := \{\eta \in L^2(\Omega) : a \leq \eta \leq b \text{ a.e. in } \Omega\}.
\end{aligned}$$

Or equivalently, define the reduced objective functional $\hat{J} : H_0^1(\Omega) \to \mathbb{R}$ by $\hat{J}(\eta) = J(S\eta, \eta)$ and consider the optimisation problem

$$\min \hat{J}(\eta) \text{ over } U_{ad}. \tag{3.8}$$

Here $g_\Gamma \in L^2(\Gamma)$, $a, b \in \mathbb{R}$ with either $a < b$ or $b = -a = \infty$, and $\nu > 0$.

Theorem 3.4. *Problem (3.8) has a unique solution $u \in U_{ad}$.*

Proof. This follows using the same argument as in Theorem 2.5, which can be found in [Tröltzsch, 2010]. □

Theorem 3.5. *$u \in L^2(\Omega)$ is a solution of (3.8) if and only if there exist $p \in H_0^1(\Omega)$ such that*

$$u \in U_{ad}, \quad (p + \nu u, v - u)_{L^2(\Omega)} \geq 0 \qquad \forall v \in U_{ad}, \tag{3.9a}$$

$$a(p, v) = \int_\Gamma (Su - g_\Gamma) v \mathrm{d}A \qquad \forall v \in H_0^1(\Omega). \tag{3.9b}$$

Proof. $\hat{J} : L^2(\Omega) \to \mathbb{R}$ has a Gâteaux derivative $\hat{J}' : L^2(\Omega) \to L^2(\Omega)^*$ and is (strictly) convex, so u is a solution of (3.8) iff

$$u \in U_{ad}, \quad \langle \hat{J}'(u), v - u \rangle_{L^2(\Omega)^*} \geq 0 \quad \forall v \in U_{ad}. \tag{3.10}$$

Calculating $\hat{J}'(u)$ we see that (3.10) becomes

$$u \in U_{ad}, \quad \int_\Gamma (Su - g_\Gamma) S(v - u) \mathrm{d}A + \nu (u, v - u)_{L^2(\Omega)} \geq 0 \quad \forall v \in U_{ad}. \tag{3.11}$$

Let $f_u(v) := \int_\Gamma (Su - g_\Gamma) v$, so $f_u \in H^{-1}(\Omega)$. Take $p = S^* f_u$, then by Lemma 3.1 we have $p \in H_0^1(\Omega)$ and it satisfies (3.9b). Also

$$(p, v-u)_{L^2(\Omega)} = (S^* f_u, v-u)_{L^2(\Omega)} = \langle f_u, S(v-u) \rangle_{H^{-1}(\Omega)} = \int_\Gamma (Su - g_\Gamma) S(v-u) \mathrm{d}A.$$

Therefore (3.11) is equivalent to (3.9a), which proves the result. □

Corollary 3.6. *If $u \in U_{ad}$ is a solution of (3.8) then it has the additional regularity that $u \in H_0^1(\Omega)$.*

Proof. This follows because $p \in H_0^1(\Omega)$, so u gains additional regularity from it through (3.9a), as in Corollary 2.7. □

Remark 3.7. *By increasing the relative weight given to the fidelity term this problem could be related to an optimal control problem with state constraints. We do not include any results on this, but a result similar to Theorem 2.8 could be proved.*

3.3 Discretisation

In this section we will formulate a discrete problem that is related to ($\mathrm{M2}_h$) from the previous chapter (see (2.34)) i.e. a discrete problem that uses the variational discretisation idea of [Hinze, 2005]. We only consider this discretisation, and not one related to ($\mathrm{M1}_h$), because we were able to prove better a priori error estimates for it in the case of control constraints.

In order to define this discrete problem in the previous chapter we simply replaced the control-to-state operator with a discrete control-to-state operator. After doing this for our current problem we are still left with a discrete problem that may be hard to solve computationally. For example, if Γ has a complicated form then it may be difficult to calculate integrals of functions defined over Γ, making the implementation of a standard finite element method impractical. Therefore we will allow in our discretisation the approximation of Γ with another hypersurface Γ_σ. If chosen carefully this may simplify the calculations needed for a numerical method, but still allow us to prove the same error estimates. In particular, we want to consider taking Γ_σ to be a polyhedral interpolation of Γ. In order to formulate such a discrete problem we now make stronger assumptions on Ω, A and Γ than were necessary to pose the continuous problem (3.8).

From now onwards suppose that Ω is convex. This simplifies the presentation by ensuring that the finite element space for the state (defined shortly) is a subset of $C_0(\Omega)$. Also from now onwards assume that the boundary of Ω and the coefficient functions a_{ij} and a_0 in the elliptic operator A are sufficiently smooth that for $2 \leq s < \frac{2n}{n-2}$,

$$\|S\eta\|_{W^{2,s}(\Omega)} \leq C\|\eta\|_{L^s(\Omega)} \quad \forall \eta \in L^s(\Omega) \tag{3.12}$$

$$\|S\eta\|_{W^{1,\infty}(\Omega)} \leq C\|\eta\|_{H^1(\Omega)} \quad \forall \eta \in H^1(\Omega). \tag{3.13}$$

Here and throughout this paper C is a positive constant that may vary from line to line and is independent of the variables it precedes. This holds, for example, when $A = -\Delta$ and Ω is smooth (see e.g. Theorem 9.9 in [Gilbarg and Trudinger, 2001] and Theorem 5 in Section 6.6 in [Evans, 2010]).

In addition to the assumptions in Section 3.1, suppose that Γ is orientable. This means that Γ has a unit normal vector field μ that is continuous (see Example 3.2), allowing us to construct the one sided strip

$$U_\delta := \{c + D\mu(c) \in \mathbb{R}^n : c \in \Gamma, -\delta < D < \delta\}.$$

Since Γ is C^2 there exists a $\delta > 0$ such that for each $x \in U_\delta$ there is a unique $c(x) \in \Gamma$ and some $-\delta < D(x) < \delta$ satisfying

$$x = c(x) + D(x)\mu(c(x)) \tag{3.14}$$

(see Lemma 2.8 [Dziuk and Elliott, 2013]). We call $D(x) : U_\delta \to \mathbb{R}$ a signed distance function for Γ, and it makes sense for both open and closed Γ. When Γ is closed it

agrees with the usual definition of the signed distance function d on U_δ (see e.g. page 296 in [Dziuk and Elliott, 2013]). Therefore all the results we need from [Dziuk and Elliott, 2013] that are proved using d also hold for our possibly open hypersurfaces using D.

Let Γ_σ be a family of Lipschitz hypersurfaces contained in $U_\delta \cap \Omega$ that are indexed by the parameter $\sigma > 0$. We intend this family of hypersurfaces to increasingly well approximate Γ as $\sigma \to 0$. We suppose that they satisfy a covering condition for each connected component Γ; for each $c \in \Gamma$ there is a unique $x \in \Gamma_\sigma$ with $c = c(x)$, where $c(x)$ is defined by (3.14). Two possible constructions of Γ_σ that we will later consider are:

- $\Gamma_\sigma := \Gamma$ for all $\sigma > 0$ i.e. we do not approximate Γ;
- Γ_σ is the union of finitely many closed $(n-1)$-simplices with maximum diameter σ. In particular, we will suppose Γ_σ is a polygonal or polyhedral interpolation of Γ. Note that such Γ_σ will always violate the covering condition when $n = 3$ unless Γ has a polygonal boundary.

Since the Γ_σ are Lipschitz we can define the function spaces $L^2(\Gamma_\sigma)$ and $C(\Gamma_\sigma)$ in the same way as for Γ.

As in Section 2.3 we can approximate Ω with a family of interpolating polygonal or polyhedral approximations Ω_h such that $|\Omega \setminus \Omega_h| \leq Ch^2$. We then define a conforming, shape regular triangulation T_h of Ω_h and the usual family of discrete spaces of piecewise linear finite elements which vanish on the boundary:

$$V_h := \{v_h \in C_0(\Omega) : v_h|_T \in P_1(T) \text{ for all } T \in T_h \text{ and } v_h|_{\Omega \setminus \Omega_h} = 0\}.$$

We use this to define a discrete approximation to S. For $\eta \in L^2(\Omega)$ let y_h be the unique function in satisfying

$$y_h \in V_h, \quad a(y_h, v_h) = (\eta, v_h) \quad \forall v_h \in V_h,$$

and define $S_h : L^2(\Omega) \to V_h$ by $S_h\eta = y_h$. Observe that V_h is a finite dimensional subspace of $H_0^1(\Omega)$, so it a Banach space when equipped with the $H_0^1(\Omega)$ norm. Therefore S_h is a linear and continuous operator between Banach spaces and we are able to define an adjoint operator.

Remark 3.8. *We choose the range of S_h to be $V_h \subset C(\bar{\Omega})$ rather than the range of S (as in Chapter 2) so that $S_h\eta|_{\Gamma_\sigma}$ is well defined and belongs to $L^2(\Gamma_\sigma)$; Lemma 3.3 does not apply since we assume that Γ_σ is Lipschitz rather than C^2.*

For $\eta \in L^2(\Omega)$ we have by (3.12) and a Sobolev embedding result that $S\eta \in C(\bar{\Omega})$, so it makes sense to look at $\|S\eta - S_h\eta\|_\infty$, where $\|\cdot\|_\infty$ denotes the supremum norm. Recall that we proved in Corollary 2.10 that for $\eta \in H^1(\Omega)$ and any $\varepsilon > 0$,

$$\|S\eta - S_h\eta\|_\infty \leq \begin{cases} C(\varepsilon)h^{2-\varepsilon}\|\eta\|_{H^1(\Omega)} & n = 2, \\ Ch^{\frac{3}{2}}\|\eta\|_{H^1(\Omega)} & n = 3. \end{cases} \tag{3.15}$$

We are now ready to introduce the discrete problem. Define the discrete reduced objective functional $\hat{J}_h : L^2(\Omega) \to \mathbb{R}$ by

$$\hat{J}_h(\eta) := \frac{1}{2}\|S_h\eta - g_{\Gamma,\sigma}\|^2_{\Gamma_\sigma} + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)}$$

and consider the following discrete problem based on the variational discretisation concept from [Hinze, 2005]:

$$\min \hat{J}_h(\eta) \text{ over } \eta \in U_{ad}. \tag{3.16}$$

Here $g_{\Gamma,\sigma} \in L^2(\Gamma_\sigma)$ is a function that will be defined to approximate $g_\Gamma \in L^2(\Gamma)$. Also let the norm $\|\cdot\|_{\Gamma_\sigma} := \sqrt{m_\sigma(\cdot,\cdot)}$, where $m_\sigma : L^2(\Gamma_\sigma) \times L^2(\Gamma_\sigma) \to \mathbb{R}$ is some inner product that will be defined to approximate the $L^2(\Gamma)$ inner product. Note that the restriction of $S_h\eta$ to $L^2(\Gamma_\sigma)$ is well defined by Remark 3.8. The assumptions we have made so far on Γ_σ, $g_{\Gamma,\sigma}$ and m_σ are sufficient to prove existence of a solution to (3.16) and derive optimality conditions. These solutions will not necessarily closely approximate the solution of the continuous problem (3.8), but we will impose further assumptions in the next section which ensure this.

Theorem 3.9. *Problem (3.16) has a unique solution $u_h \in U_{ad}$. Moreover, $u_h \in L^2(\Omega)$ is a solution of (3.16) if and only if there exists $p_h \in V_h$ such that*

$$u_h \in U_{ad}, \quad (p_h + \nu u_h, v - u_h) \geq 0 \qquad \forall v \in U_{ad} \tag{3.17a}$$

$$a(v_h, p_h) = m_\sigma(S_h u_h|_{\Gamma_\sigma} - g_{\Gamma,\sigma}, v_h|_{\Gamma_\sigma}) \qquad \forall v_h \in V_h. \tag{3.17b}$$

Proof. The proof follows by the same arguments as in Theorems 2.5 and 3.5. □

We are minimising over the infinite dimensional space U_{ad}, but (3.17a) implies

$$u_h = \mathbb{P}_{[a,b]}\Big(-\frac{1}{\nu}p_h\Big) \quad \text{a.e. in } \Omega.$$

So the control is implicitly discretised through S_h, as it was in ($\mathrm{M2}_h$) (see (2.34)). This means that the above optimality conditions can be solved computationally for

appropriate choices of Γ_σ, m_σ and $g_{\Gamma,\sigma}$.

3.4 Numerical analysis

In this section we will prove an a priori $L^2(\Omega)$ error estimate for convergence of the discrete optimal control problem (3.16) to the continuous optimal control problem (3.8). This will require additional assumptions on Γ_σ, m_σ and $g_{\Gamma,\sigma}$. In order to write down these assumptions we need a way to compare functions defined on Γ_σ with functions defined on Γ. For this purpose we introduce the lift operator (see e.g. Section 4.1 in [Dziuk and Elliott, 2013] for more details).

Let w_σ be a function defined on Γ_σ. Due to the covering condition, for each $c \in \Gamma$ there is a unique $x \in \Gamma_\sigma$ with $c = c(x)$ (see (3.14). We will denote this x by $x(c)$. Then the lift operator $(\cdot)^l$ mapping a function defined on Γ_σ to a function defined on Γ is given by

$$w_\sigma^l(c) := w_\sigma(x(c)) \quad \forall c \in \Gamma.$$

Note that the inverse lift operator $(\cdot)^{-l} := ((\cdot)^l)^{-1}$ is well defined. We also use $x(c)$ to define the distance $D_\sigma : \Gamma \mapsto \mathbb{R}$ between Γ and Γ_σ by

$$D_\sigma(c) := |c - x(c)| .$$

We can now impose the following additional assumptions on Γ_σ, m_σ and $g_{\Gamma,\sigma}$ (see Section 3.3 for the previous assumptions).

Assumption 3.10. *Γ_σ, m_σ and $g_{\Gamma,\sigma}$ satisfy*

$$\sup_{c\in\Gamma} D_\sigma(c) \leq C\sigma^2, \tag{3.18}$$

$$\left|(w_\sigma^l, z_\sigma^l)_{L^2(\Gamma)} - m_\sigma(w_\sigma, z_\sigma)\right| \leq C\sigma^2 \|w_\sigma^l\|_{L^2(\Gamma)} \|z_\sigma^l\|_{L^2(\Gamma)} \quad \forall w_\sigma, z_\sigma \in L^2(\Gamma_\sigma), \tag{3.19}$$

$$\|g_\Gamma - g_{\Gamma,\sigma}^l\|_{L^2(\Gamma)} \leq C\sigma^2 \tag{3.20}$$

with C independent of σ.

Under these assumptions we can prove some lemmas which will enable us to prove a priori $L^2(\Omega)$ error estimates for the control.

Lemma 3.11. *Let u_h be the solution of (3.16). For sufficiently small σ and h,*

$$\|u_h\|_{H^1(\Omega)} + \|p_h\|_{H^1(\Omega)} \leq Ch^{-\frac{n}{2}}\sigma^2$$

with C independent of σ and h.

Proof. Fix $\sigma_0,\ h_0 \in \mathbb{R}$ and suppose throughout this proof that $0 < \sigma < \sigma_0$ and $0 < h < h_0$.

First observe that $\|u_h\|_{H_0^1(\Omega)} \leq C\|p_h\|_{H_0^1(\Omega)}$, since (3.17a) gives that $u_h = \mathbb{P}_{[a,b]}\left(-\frac{1}{\nu}p_h\right)$. Also, for $v \in H_0^1(\Omega)$ we have by the Poincaré inequality that $\|v\|_{H^1(\Omega)} \leq C\|v\|_{H_0^1(\Omega)}$. So we just need to show that $\|p_h\|_{H_0^1(\Omega)} \leq C$.

Testing (3.17b) with $v_h = p_h$ and using the coercivity of $a(\cdot,\cdot)$ and the boundedness of $m_\sigma(\cdot,\cdot)$ we get that

$$\begin{aligned}
\alpha\|p_h\|_{H_0^1(\Omega)}^2 &\leq C\|S_h u_h - g_{\Gamma,\sigma}\|_{L^2(\Gamma_\sigma)}\|p_h\|_{L^2(\Gamma_\sigma)} \\
&\leq C(\|S_h u_h\|_{L^2(\Gamma_\sigma)} + \|g_{\Gamma,\sigma}^l\|_{L^2(\Gamma)})\|p_h^l\|_{L^2(\Gamma)} \\
&\leq C(\|S_h u_h\|_{L^\infty(\Omega)} + 1)(\|p_h\|_{L^2(\Gamma)} + \|p_h^l - p_h\|_{L^2(\Gamma)}) \\
&\leq C(\|u_h\|_{L^2(\Omega)} + 1)(\|p_h\|_{H_0^1(\Omega)} + \|p_h^l - p_h\|_{L^2(\Gamma)}) \\
&\leq C(\|p_h\|_{H_0^1(\Omega)} + \|p_h^l - p_h\|_{L^2(\Gamma)}). \qquad (3.21)
\end{aligned}$$

For the third inequality we have used assumption (3.20). For the fourth inequality have used the supremum norm error estimate (2.26) and the trace inequality from Lemma 3.3. For the last inequality we have used that

$$\frac{\nu}{2}\|u_h\|_{L^2(\Omega)}^2 \leq \hat{J}_h(u_h) = \hat{J}_h(0) = \|g_{\Gamma,\sigma}\|_{\Gamma_\sigma}^2 \leq \|g_{\Gamma,\sigma}^l\|_{L^2(\Gamma)}^2 \leq C,$$

which implies $\|u_h\|_{L^2(\Omega)}^2 \leq C$.

Note that for $v \in W^{1,\infty}(\Omega)$ and $x_1, x_2 \in \Omega$,

$$|v(x_1) - v(x_2)| \leq \|\nabla v\|_{L^\infty(\Omega)}\,|x_1 - x_2|\,,$$

(see Theorem 2.1.4 in [Ziemer, 1989]) so

$$\begin{aligned}
\|v^l - v\|_{L^2(\Gamma)} &= \left(\int_\Gamma \big(v(c) - v(x(c))\big)^2 \mathrm{d}c\right)^{\frac{1}{2}} \\
&\leq \|\nabla v\|_{L^\infty(\Omega)} \sup_{c\in\Gamma} |c - x(c)| \\
&\leq \|v\|_{W^{1,\infty}(\Omega)} \sup_{c\in\Gamma} D_\sigma(c). \qquad (3.22)
\end{aligned}$$

Using this with $v = p_h$, an inverse inequality, and assumption (3.18) we get

$$\|p_h^l - p_h\|_{L^2(\Gamma)} \leq Ch^{-\frac{n}{2}}\sigma^2\|p_h\|_{H_0^1(\Omega)}.$$

Combining this with (3.21) gives

$$\alpha\|p_h\|^2_{H^1_0(\Omega)} \leq C\|p_h\|_{H^1_0(\Omega)}(1 + h^{-\frac{n}{2}}\sigma^2),$$

so the result follows. □

Lemma 3.12. *For some $\eta \in H^1_0(\Omega)$ set $w := S\eta|_\Gamma - g_\Gamma$ and $w_\sigma := S_h\eta|_{\Gamma_\sigma} - g_{\Gamma,\sigma}$. Then for sufficiently small σ and h,*

$$\left|\|w\|^2_{L^2(\Gamma)} - \|w_\sigma\|^2_{\Gamma_\sigma}\right| \leq C(\|\eta\|_{H^1(\Omega)})\left(\|S\eta - S_h\eta\|_\infty + \sigma^2\right)$$

with C independent of σ and h.

Proof. Fix σ_0, $h_0 \in \mathbb{R}$ and suppose throughout this proof that $0 < \sigma < \sigma_0$ and $0 < h < h_0$.

Make the splitting

$$\left|\|w\|^2_{L^2(\Gamma)} - \|w_\sigma\|^2_{\Gamma_\sigma}\right| \leq \left|\|w\|^2_{L^2(\Gamma)} - \|w^l_\sigma\|^2_{L^2(\Gamma)}\right| + \left|\|w^l_\sigma\|^2_{L^2(\Gamma)} - \|w_\sigma\|^2_{\Gamma_\sigma}\right|.$$

To bound the first term on the right hand side note that

$$\begin{aligned}\left|\|w\|^2_{L^2(\Gamma)} - \|w^l_\sigma\|^2_{L^2(\Gamma)}\right| &= \left|(w + w^l_\sigma, w - w^l_\sigma)\right| \\ &\leq \|w + w^l_\sigma\|_{L^2(\Gamma)}\|w - w^l_\sigma\|_{L^2(\Gamma)} \\ &\leq \left(\|w\|_{L^2(\Gamma)} + \|w^l_\sigma\|_{L^2(\Gamma)}\right)\|w - w^l_\sigma\|_{L^2(\Gamma)}. \end{aligned} \tag{3.23}$$

Using the trace result from Lemma 3.3 and the continuity of S we have

$$\begin{aligned}\|w\|_{L^2(\Gamma)} &= \|S\eta - g_\Gamma\|_{L^2(\Gamma)} \\ &\leq \|S\eta\|_{L^2(\Gamma)} + \|g_\Gamma\|_{L^2(\Gamma)} \\ &\leq C\|\eta\|_{L^2(\Omega)} + \|g_\Gamma\|_{L^2(\Gamma)} \\ &\leq C(\|\eta\|_{L^2(\Omega)}).\end{aligned}$$

Similarly using assumption (3.20) and the supremum norm error estimate (2.26) we

get

$$\begin{aligned}\|w_\sigma^l\|_{L^2(\Gamma)} &= \|S_h\eta - g_{\Gamma,\sigma}\|_{L^2(\Gamma_\sigma)} \\ &\leq \|S_h\eta\|_{L^2(\Gamma_\sigma)} + \|g_{\Gamma,\sigma}^l\|_{L^2(\Gamma)} \\ &\leq C(\|S_h\eta\|_\infty + 1) \\ &\leq C(\|\eta\|_{L^2(\Omega)}). \end{aligned} \tag{3.24}$$

Using (3.13), assumption (3.20) and the estimate (3.22) we get

$$\begin{aligned}\|w - w_\sigma^l\|_{L^2(\Gamma)} &\leq \|S\eta - (S_h\eta)^l\|_{L^2(\Gamma)} + \|g_\Gamma - g_{\Gamma,\sigma}^l\|_{L^2(\Gamma)} \\ &\leq \|S\eta - (S\eta)^l\|_{L^2(\Gamma)} + \|(S\eta)^l - (S_h\eta)^l\|_{L^2(\Gamma)} + C\sigma^2 \\ &\leq C(\sigma^2\|S\eta\|_{W^{1,\infty}(\Omega)} + \|S\eta - S_h\eta\|_\infty + \sigma^2) \\ &\leq C(\sigma^2\|\eta\|_{H^1(\Omega)} + \|S\eta - S_h\eta\|_\infty + \sigma^2) \\ &\leq C(\|\eta\|_{H^1(\Omega)})(\|S\eta - S_h\eta\|_\infty + \sigma^2).\end{aligned}$$

Combining these estimates with (3.23), the bound for the first term in the splitting becomes

$$\left|\|w\|_{L^2(\Gamma)}^2 - \|w_\sigma^l\|_{L^2(\Gamma)}^2\right| \leq C(\|\eta\|_{H^1(\Omega)})(\|S\eta - S_h\eta\|_\infty + \sigma^2).$$

We can bound the second term in the splitting using assumption (3.19) and the estimate (3.24). This completes the proof. □

Remark 3.13. *It is now clear why our assumptions involve σ^2 bounds as opposed to some other power. When $\sigma = h$, this rate of convergence will not dominate the $h^{2-\varepsilon}$ supremum norm error estimate (2.28) (for $n = 2$), which we will use to bound $\|S\eta - S_h\eta\|_\infty$ on the right hand side of Lemma 3.12.*

Note that if we take $\Gamma_\sigma = \Gamma$, $m_\sigma = m$ and $g_{\Gamma,\sigma} = g_\Gamma$ then the assumptions are trivially satisfied. We will later see that there are nontrivial definitions based on polyhedral interpolations of Γ that satisfy the assumptions. If we use polyhedral approximations of Γ that are not interpolating then these assumptions may not be satisfied.

We are ready to use the approach of [Deckelnick and Hinze, 2007; Leykekhman et al., 2013], as in Section 2.4.2, to prove the following error estimate.

Theorem 3.14. *Suppose Assumption 3.10 holds. Let u and u_h be solutions of (3.8)*

and (3.16) respectively. If $\sigma \leq Ch^{\frac{n}{4}}$ and h is sufficiently small then for any $\varepsilon > 0$,

$$\|u - u_h\|_{L^2(\Omega)} \leq C\left(\sigma + \begin{cases} C(\varepsilon)h^{1-\varepsilon} & n = 2, \\ h^{\frac{3}{4}} & n = 3 \end{cases}\right)$$

with C independent of σ and h.

Proof. Fix σ_0, $h_0 \in \mathbb{R}$ and suppose throughout this proof that $0 < \sigma < \sigma_0$ and $0 < h < h_0$.

Rearranging and using the optimality conditions as in the proof of Theorem 2.25 we see that

$$\begin{aligned} \frac{1}{2}\|Su - Su_h\|^2_{L^2(\Gamma)} + \frac{\nu}{2}\|u - u_h\|^2_{L^2(\Omega)} &\leq \hat{J}(u_h) - \hat{J}(u), \\ \frac{1}{2}\|S_h u - S_h u_h\|^2_{L^2(\Gamma_\sigma)} + \frac{\nu}{2}\|u - u_h\|^2_{L^2(\Omega)} &\leq \hat{J}_h(u) - \hat{J}_h(u_h). \end{aligned}$$

Adding these two relations we get

$$\begin{aligned} \nu\|u - u_h\|^2_{L^2(\Omega)} &\leq \hat{J}(u_h) - \hat{J}(u) - \hat{J}_h(u_h) + \hat{J}_h(u) \\ &\leq \left|\hat{J}(u) - \hat{J}_h(u)\right| + \left|\hat{J}(u_h) - \hat{J}_h(u_h)\right|. \end{aligned} \tag{3.25}$$

Lemma 3.12 gives the estimate

$$\begin{aligned} \left|\hat{J}(u) - \hat{J}_h(u)\right| &= \left|\|Su - g_\Gamma\|^2_{L^2(\Gamma)} - \|S_h u - g_{\Gamma,\sigma}\|^2_{\Gamma_\sigma}\right| \\ &\leq C(\|u\|_{H^1(\Omega)})\left(\|Su - S_h u\|_\infty + \sigma^2\right) \\ &\leq C\left(\|Su - S_h u\|_\infty + \sigma^2\right) \end{aligned}$$

with C independent of σ and h. Now using the supremum norm error estimate (2.28) we get that for any $\varepsilon > 0$,

$$\left|\hat{J}(u) - \hat{J}_h(u)\right| \leq C(\|u\|_{H^1(\Omega)})\left(\sigma^2 + \begin{cases} C(\varepsilon)h^{2-\varepsilon}\|u\|_{H^1(\Omega)} & n = 2, \\ h^{\frac{3}{2}}\|u\|_{H^1(\Omega)} & n = 3 \end{cases}\right). \tag{3.26}$$

The same approach gives the estimate

$$\left|\hat{J}(u_h) - \hat{J}_h(u_h)\right| \leq C(\|u_h\|_{H^1(\Omega)})\left(\sigma^2 + \begin{cases} C(\varepsilon)h^{2-\varepsilon}\|u_h\|_{H^1(\Omega)} & n = 2, \\ h^{\frac{3}{2}}\|u_h\|_{H^1(\Omega)} & n = 3 \end{cases}\right), \tag{3.27}$$

where the $\|u_h\|_{H^1(\Omega)}$ term comes from using the supremum norm error estimate

(2.28). If we take $\sigma \leq Ch^{\frac{n}{4}}$ then by Lemma 3.11 we can bound $\|u_h\|_{H^1(\Omega)}$ independently of h.

Combining (3.26) and (3.27) with (3.25) completes the proof. □

Remark 3.15. *We can compare this error estimate to those for analogous discretisations of the standard optimal control problem and the point optimal control problem of the previous chapter.*

- *Standard control problem (1.6) discretised as in [Hinze, 2005]:*

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^2.$$

- *Point control problem (2.12) discretised by (2.34): For any $\varepsilon > 0$,*

$$\|u - u_h\|_{L^2(\Omega)} \leq C \begin{cases} h^{1-\varepsilon} & n = 2, \\ h^{\frac{1}{2}-\varepsilon} & n = 3. \end{cases}$$

- *Control on surface problem (3.8) discretised by (3.16): With $\sigma = h$, for any $\varepsilon > 0$,*

$$\|u - u_h\|_{L^2(\Omega)} \leq \begin{cases} C(\varepsilon)h^{1-\varepsilon} & n = 2, \\ Ch^{\frac{3}{4}} & n = 3. \end{cases}$$

3.4.1 Example definitions of Γ_σ, m_σ and $g_{\Gamma,\sigma}$

So far we have just stated properties that Γ_σ, m_σ and $g_{\Gamma,\sigma}$ must have in order for Theorem 3.14 to hold for the discrete problem (3.8). We now give some definitions for these quantities that satisfy all the required properties. Different definitions will lead to discrete problems that are easier or harder to solve, and so the definitions we use in practice will depend on Γ and g_Γ.

Method 1

Take the following definitions for Γ_σ, m_σ and $g_{\Gamma,\sigma}$ in the discrete problem (3.8):

- $\Gamma_\sigma := \Gamma$ i.e. do not approximate Γ.
- $m_\sigma(w_\sigma, z_\sigma) := \int_\Gamma w_\sigma z_\sigma \mathrm{d}A$. This trivially satisfies assumption (3.19) since $w^l = w$ for $w \in L^2(\Gamma)$.
- $g_{\Gamma,\sigma} := g_\sigma$. This trivially satisfies assumption (3.20).

Theorem 3.14 holds since all the assumptions are satisfied.

We would typically take these choices when Γ and g_Γ have simple forms. For example, perhaps when Γ is a straight line or circle and g_Γ is piecewise constant function. In this case the integrals over Γ of products of discrete functions and g_Γ may be easy to compute. This would allow us to implement the numerical method described in Section 3.5 exactly.

Remark 3.16. *In practice computing the required integrals over Γ will be difficult, even when Γ and g_Γ are simple. One way to handle this is to use a quadrature in the implementation. See Section 3.4.2 for a related discussion.*

Method 2

Suppose $g_\Gamma \in H^2(\Gamma)$ and take the following definitions for Γ_σ, m_σ and $g_{\Gamma,\sigma}$ in the discrete problem (3.8):

- Let each Γ_σ consist of a union of finitely many closed $(n-1)$-simplices whose vertices lie on Γ and form a conforming, shape regular triangulation E_σ of size σ. By this we mean that $\sigma = \max_{E\in E_\sigma}\sigma(E)$ and for each element $E \in E_\sigma$ the quantity
$$\max_{E\in E_\sigma} \kappa(E), \quad \kappa(E) := \frac{\sigma(E)}{\rho(E)}$$
is uniformly bounded independently of σ. Here $\sigma(E)$ denotes the diameter of E and $\rho(E)$ denotes the diameter of the largest ball contained in E.

  Let δ_σ denote the quotient between the smooth and discrete surface measures $\mathrm{d}A$ on Γ and $\mathrm{d}A_\sigma$ on Γ_σ i.e. δ_σ is defined by $\delta_\sigma \mathrm{d}A_\sigma = \mathrm{d}A$ and
$$\int_{\Gamma_\sigma} w_\sigma \mathrm{d}A_\sigma = \int_\Gamma w_\sigma^l \frac{1}{\delta_\sigma}\mathrm{d}A \quad \forall w_\sigma \in L^2(\Gamma_\sigma). \tag{3.28}$$

  For Γ_σ defined as above, Lemma 4.1 in [Dziuk and Elliott, 2013] gives
$$\|1 - \frac{1}{\delta_\sigma}\|_{L^\infty(\Gamma)} \leq C\sigma^2, \tag{3.29}$$
and Lemma 4.2 in [Dziuk and Elliott, 2013] gives
$$\|w_\sigma^l\|_{L^2(\Gamma)} \leq C\|w_\sigma\|_{L^2(\Gamma_\sigma)} \quad \forall w_\sigma \in L^2(\Gamma_\sigma)$$
with C independent of σ and w_σ.

- $m_\sigma(w_\sigma, z_\sigma) := \int_{\Gamma_\sigma} w_\sigma z_\sigma \mathrm{d}A_\sigma$. Assumption (3.19) holds, since (3.28) and (3.29) give that for $w_\sigma, z_\sigma \in L^2(\Gamma_\sigma)$,

$$\begin{aligned}\left|(w_\sigma^l, z_\sigma^l) - m_\sigma(w_\sigma, z_\sigma)\right| &= \left|\int_\Gamma w_\sigma^l z_\sigma^l \Big(1 - \frac{1}{\delta_\sigma}\Big)\mathrm{d}A\right| \\ &\leq \|1 - \frac{1}{\delta_\sigma}\|_{L^\infty(\Gamma)} \|w_\sigma^l\|_{L^2(\Gamma)} \|z_\sigma^l\|_{L^2(\Gamma)} \\ &\leq C\sigma^2 \|w_\sigma^l\|_{L^2(\Gamma)} \|z_\sigma^l\|_{L^2(\Gamma)}.\end{aligned}$$

- $g_{\Gamma,\sigma} := I_\sigma g_\Gamma$, where I_σ is the Lagrange interpolation of $g_\Gamma \in H^2(\Gamma)$ onto

$$W_\sigma := \{w_\sigma \in C(\Gamma_\sigma) : w_\sigma|_E \in P_1(E) \text{ for all } E \in E_\sigma\},$$

  the space of piecewise affine finite elements on E_σ. In particular, $I_\sigma(w) := (\tilde{I}_\sigma w^{-l})$ where $\tilde{I}_\sigma : C(\Gamma_\sigma) \to W_\sigma$ is the Lagrange interpolation operator. By $I_\sigma^l w$ denote $(I_\sigma w)^l$, then for $w \in H^2(\Gamma) \subset C(\Gamma)$ we have

$$\|w - I_\sigma^l w\|_{L^2(\Gamma)} \leq C\sigma^2 \|w\|_{H^2(\Gamma)}$$

  (see [Dziuk, 1988; Demlow, 2009]). So assumption (3.20) is satisfied if $g_\Gamma \in H^2(\Gamma)$.

Since all the assumptions are satisfied, Theorem 3.14 holds.

We may want to use these definitions of Γ_σ, m_σ and $g_{\Gamma,\sigma}$ if Γ has a complicated form. In this case it is likely to be hard to calculate integrals over Γ, which are required by our numerical method (described in Section 3.5). By approximating Γ with a polygonal or polyhedral Γ_σ we only need to compute integrals over straight lines or triangles, which is easier. Note that even if g_Γ is quite simple, a complicated Γ means that g_Γ^l could be complicated. This is why we also define $g_{\Gamma,h}$ to be the above piecewise affine interpolation of g_Γ. Then the surface integrals that are needed for our numerical method simplify to integrals of products of piecewise linear functions over flat surfaces. These are fairly straightforward to calculate and implement.

Remark 3.17. *There are a few natural approaches to defining an interpolating polygonal or polyhedral Γ_σ (see Figure 3.1). These different approaches lead to different challenges. For our numerics we will use approach (c) in the figure, which ensures Γ_σ coincides with edges (for $n = 2$) of T_h. This simplifies the calculation of integrals over Γ_σ, but constructing a suitable T_h may be hard. It also effectively forces $\sigma = h$.*

Remark 3.18. *Theorem 3.14 says that for $n = 3$ we could take $\sigma = h^{\frac{3}{4}}$ without dominating the error from the discretisation of the state. We could make use of this if we were to instead use approach (a) in Figure 3.1.*

3.4.2 Link to optimal control at points

Method 2 can be thought of as using a quadrature to approximate Method 1. Note that

$$w_\sigma := S_h\eta|_{\Gamma_\sigma} - I_h g_\Gamma \quad \forall \eta \in L^2(\Omega)$$

is piecewise linear on Γ_σ. Therefore

$$\|w_\sigma\|^2_{\Gamma_\sigma} = \int_{\Gamma_\sigma} w_\sigma^2 \mathrm{d}A_h$$

corresponds to integrating a piecewise quadratic function on Γ_σ. This can be computed exactly with a weighted sum of point evaluations. In particular, Method 2 can be equivalently written as a discrete point control problem of a similar form to ($\mathrm{M2}_h$) (see (2.34)) from the previous chapter:

$$\min \frac{1}{2}\sum_{w\in I} \kappa_\omega \left|S_h\eta(\omega) - g_\omega\right|^2 + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)} \text{ over } \eta \in L^2(\Omega),$$

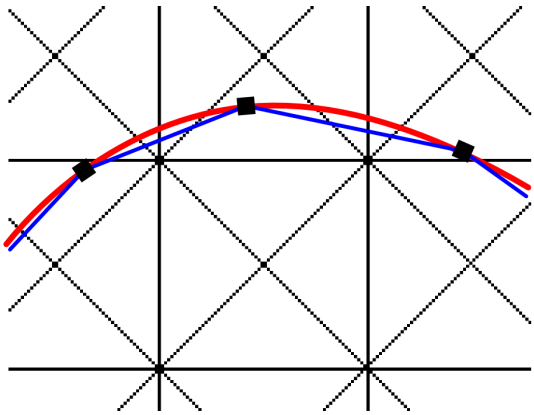
where $g_\omega := g_\Gamma(\omega)$, the set I contains points in Γ, and κ_ω are weights. If we construct a triangulation that contains Γ_σ as edges (i.e. use approach (c) in Figure 3.1) then $|I| = O(\frac{1}{h})$ and the κ_ω are $O(h)$.

As Theorem 3.14 holds using Method 2, we have provided an example of solutions to discrete point control problems that converge to the solution of a surface control problem.

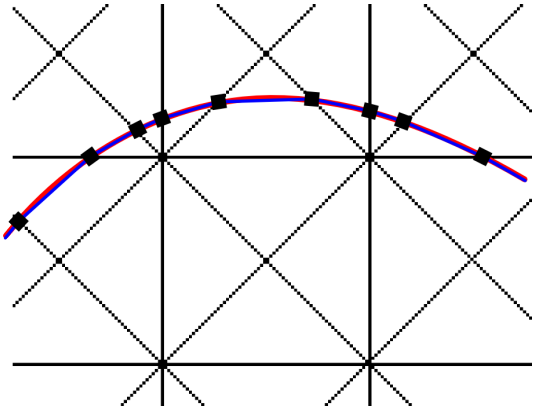
Remark 3.19. *We could also consider a weighted fidelity term for the surface control problem i.e. replace the fidelity term in (3.1) by*

$$\frac{1}{2}\int_\Gamma w(y - g_\Gamma)^2 \mathrm{d}A,$$

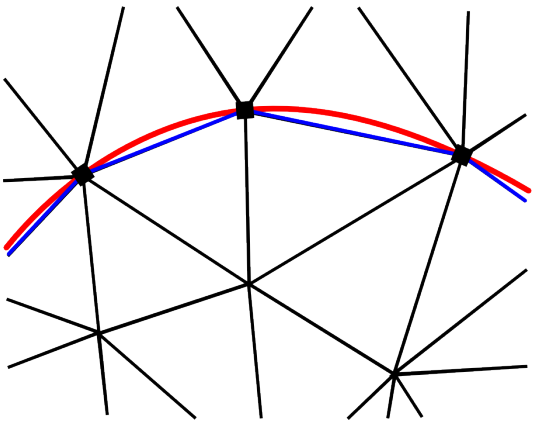
where $w \in L^\infty(\Omega)$ and $w \geq 0$. After the obvious modifications, all the results proved in this chapter would still hold.



(a) Here we take an arbitrary interpolation of Γ. This does not have any relation to the triangulation T_h, so calculating integrals over Γ_σ may be tricky.



(b) Here we construct the interpolation of Γ using the triangulation T_h; the beginning and end of segments of Γ_σ are given by the points where Γ intersects the edges of T_h. This makes calculating integrals over Γ_σ easier.



(c) Here Γ_σ is chosen first then T_h is constructed to contain the segments of Γ_σ as edges. This leads to the easiest calculation of integrals over Γ_σ, but constructing T_h may be hard.

Figure 3.1: An illustration of different constructions of T_h and polygonal Γ_σ for $n = 2$. The black lines mark the triangulation T_h, the red curve is Γ and the blue curve is Γ_σ. The square markers indicate the beginning and end of segments of Γ_σ.

3.5 Numerical results

In this section we describe the numerical method we use to solve (3.16) and show that the error estimate from Theorem 3.14 for $n = 2$ is observed in practice.

3.5.1 Numerical method

The numerical method is similar to the one described in Section 2.5.1. Analogously to (2.58) (but for a discrete problem without a forcing term f and the point evaluation term replaced by a surface integral term) we have that $y_h := S_h u_h$ and the p_h satisfying the optimality conditions (3.16) solve

$$\begin{pmatrix} a(y_h, v_h) - (-\frac{1}{\nu}p_h + (a + \frac{1}{\nu}p_h)^+ - (-\frac{1}{\nu}p_h - b)^+, v_h) \\ a(v_h, p_h) - m_\sigma(y_h|_{\Gamma_\sigma} - g_{\Gamma,\sigma}, v_h|_{\Gamma_\sigma}) \end{pmatrix} = 0 \quad \forall v_h \in V_h,$$

where v^+ denotes the nonnegative part of v i.e. $\max(0, v)$. We can then determine u_h by calculating $u_h = \mathbb{P}_{[a.b]}\big(-\frac{1}{\nu}p_h\big)$. We use a semismooth Newton method to solve the above system, but we will not describe it in detail as it follows from only minor modifications to the one in Section 2.5.1. We only implement this numerical method for $n = 2$. The implementation for $n = 3$ would be more complicated.

Depending on the example we are considering, we may either use Method 1 or Method 2 to choose Γ_σ, m_σ and $g_{\Gamma,\sigma}$. When using Method 2 we will use the approach from Figure 3.1(c) and construct the triangulation T_h from Γ_σ: We first find a polygonal curve Γ_σ with segments of length h (i.e. we take $\sigma = h$), and then use the program Triangle (see [Shewchuk]) to construct a triangulation T_h of size h that contains the segments of Γ_σ as edges.

3.5.2 Examples

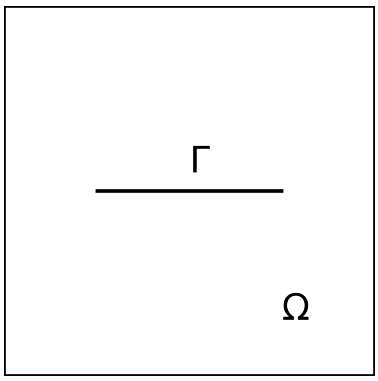
In all our examples we will take $A = -\Delta$ and $\sigma = h$. We first solve two simple examples on a Γ that is a straight line.

Example 3.20. $\Omega = (0,1)^2$, $\Gamma = \{(0.25 + 0.5t, 0.5) : t \in (0,1)\}$, $g_\Gamma(x_1, x_2) = \sin(3\pi x_1)$, $\nu = 1e-2$, $b = -a = \infty$.

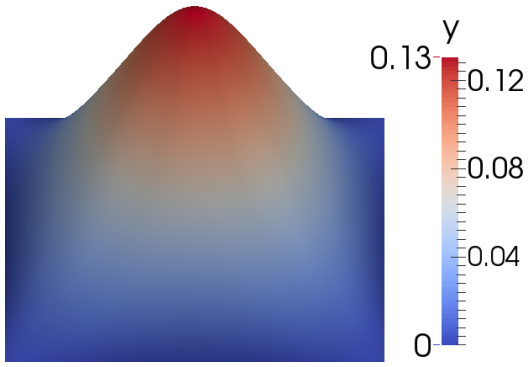
Example 3.21. *The same as Example 3.20 but with*

$$g_\Gamma(x_1, x_2) = \begin{cases} 1 & x < 0, \\ -1 & x \geq 0 \end{cases}$$

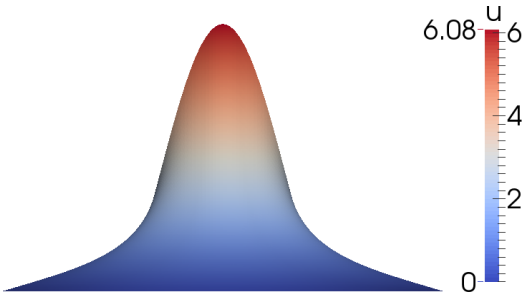
and $b = -a = 5$.



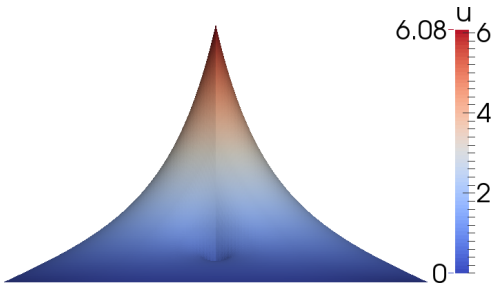
(a) Illustration of Γ.



(b) Surface plot of y_h.



(c) Surface plot of $u_h = -\frac{1}{\nu}p_h$ with the x_1-axis at the bottom.

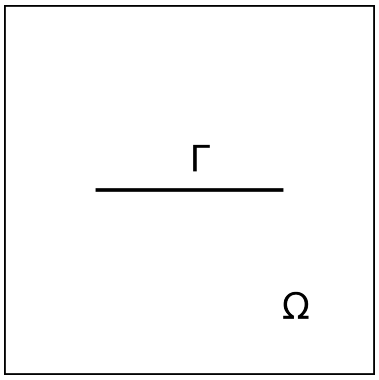


(d) Surface plot of $u_h = -\frac{1}{\nu}p_h$ with the x_2-axis at the bottom.

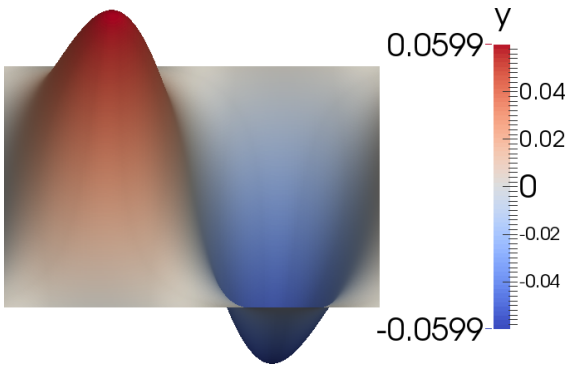
Figure 3.2: The solution to Example 3.20. We use Γ_σ, m_σ and $g_{\Gamma,\sigma}$ defined by Method 2, even though we do not need to approximate Γ, as interpolating g_Γ simplifies the implementation.

Example 3.20 has a smooth but nonconstant g_Γ and no control constraints. Its solution can be seen in Figure 3.2. Example 3.21 has a discontinuous g_Γ and active control constraints. Its solution can be seen in Figure 3.3. Even for these simple examples the exact solution is not known explicitly, so we compute L^2 errors against discrete solutions on fine triangulations to get approximate experimental orders of convergence (EOCs). In particular we use (2.61), again with $h_{\text{fine}} = 0.00276214$, which corresponds to 263169 DOFs. The approximate EOCs for these examples are in Tables 3.2 and 3.3. They agree with the error estimate we proved in Theorem 3.14 for $n = 2$. We do not verify this error estimate for examples with curved Γ: With our approach of constructing triangulations T_h that coincide with Γ_σ, the resulting T_h for a small h will not in general be a refinement of a T_h for a larger h. This makes it challenging to compute $L^2(\Omega)$ errors.

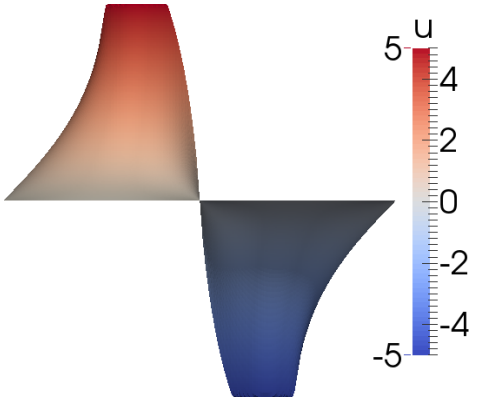
In comparison to solutions of point control problems, the solutions of these line control examples appear to have bounded p (and hence also u). An interesting feature of the solutions are the ridges in p_h and u_h along Γ. Observe that in the above



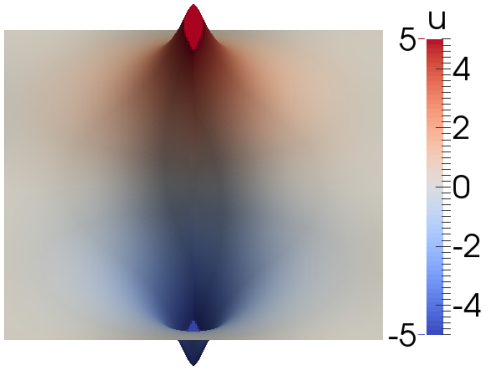
(a) Illustration of Γ.



(b) Surface plot of y_h.



(c) Surface plot of $u_h = \mathbb{P}_{[a,b]}\big(-\frac{1}{\nu}p_h\big)$ with the x_1-axis at the bottom.



(d) Surface plot of $u_h = \mathbb{P}_{[a,b]}\big(-\frac{1}{\nu}p_h\big)$ with the x_2-axis at the bottom.

Figure 3.3: The solution to Example 3.21. We use Γ_σ, m_σ and $g_{\Gamma,\sigma}$ defined by Method 1, as it is easy to integrate discrete functions against $g_{\Gamma,\sigma}$ along Γ. The figure can be interpreted in the same way as Figure 3.2.

h	#DOFs	$\|u_h - \tilde{u}\|_{L^2(\Omega)}$	EOC_h
0.353553	25	0.92096	-
0.176777	81	0.413261	1.1561
0.0883883	289	0.193377	1.0956
0.0441942	1089	0.0967977	0.9984
0.0220971	4225	0.0482398	1.0047
0.0110485	16641	0.0235625	1.0337

Table 3.2: EOCs for Example 3.20 (which has no active control constraints).

h	# DoFs	$\|u - u_h\|_{L^2(\Omega)}$	EOC_h
0.353553	25	0.991883	0
0.176777	81	0.544039	0.86646341
0.0883883	289	0.292202	0.89674110
0.0441942	1089	0.146281	0.99822529
0.0220971	4225	0.0741029	0.98114049
0.0110485	16641	0.0363584	1.0272346

Table 3.3: EOCs for Example 3.21 (which has active control constraints).

examples $y_h|_\Gamma$ does not get close to g_Γ because $\nu = 1e-2$ is too large, especially when there are control constraints. In the next examples we take $\nu = 1e-4$ and observe that we can get close agreement between $y_h|_\Gamma$ and g_Γ. In the remaining examples the only variable that will change is Γ.

Example 3.22. $\Omega = (0,1)^2$,

$$\Gamma = \{(0.5 + 0.327t\sin t, 0.5 + 0.327t\cos t) : t \in (0, 3.159)\},$$

(i.e. a spiral), $g_\Gamma = 1$, $\nu = 1e-4$, $b = -a = \infty$.

Example 3.23. *The same as Example 3.22 but with*

$$\Gamma = \{(0.5 + 0.25\cos(2\pi t), 0.5 + 0.25\sin(2\pi t) : t \in (0,1)\}$$

(i.e. a circle).

Example 3.24. *The same as Example 3.22 but with a multi-component* Γ *having the spoke like structure marked by the black lines in Figure 3.7(c).*

In Examples 3.22 and 3.23 Γ is curved. As described in Section 3.5, we first construct a Γ_σ that interpolates Γ, then create a triangulation that coincides with Γ_σ. To illustrate this a possible (but coarse) triangulation for the spiral shaped Γ from Example 3.22 is shown in Figure 3.4. In Example 3.24 the spoke like Γ is formed from a Γ consisting of multiple connected components; in particular, 6 open lines originating from the point $(0.5, 0.5)$.

Solutions to Examples 3.22, 3.23 and 3.24 can be seen in Figures 3.6, 3.7 and 3.5. These were computed with $h = 0.00292967$ and #DOFs ≈ 70000. Observe that for this small value of $\nu = 1e-4$, the values of $y_h|_\Gamma$ are close to $g_\Gamma = 1$.

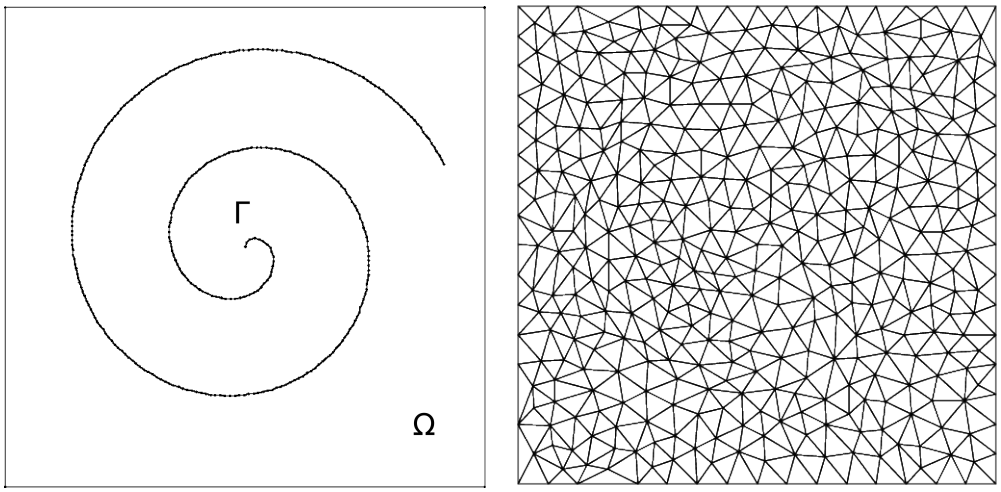
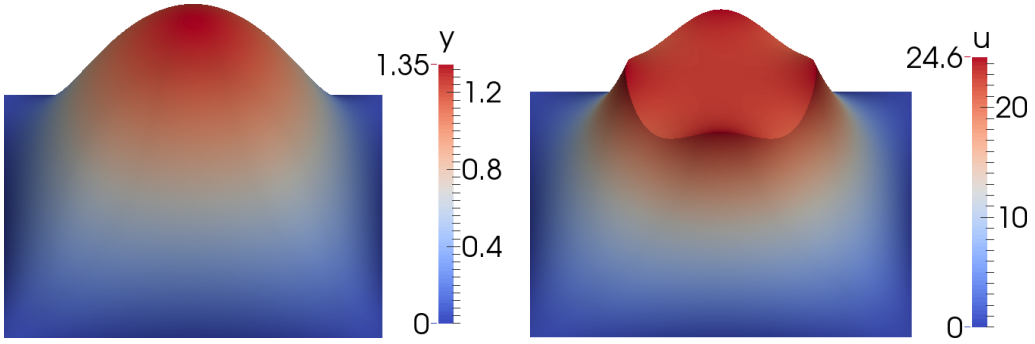
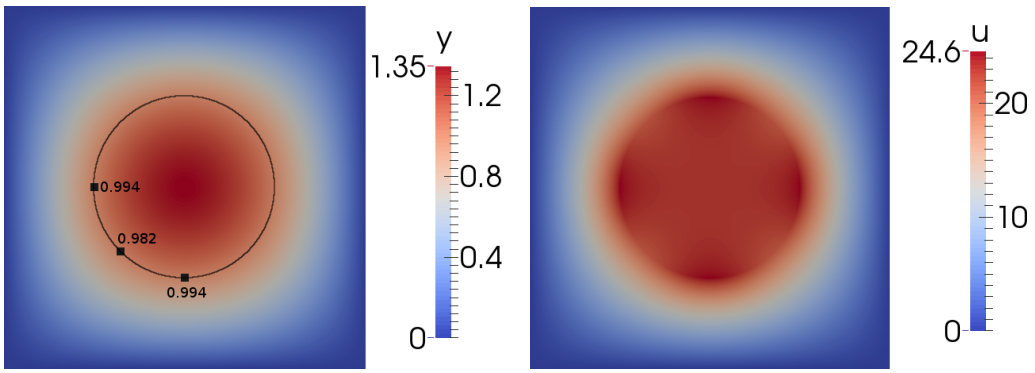


Figure 3.4: Γ as defined in Example 3.22 and a triangulation whose edges contain an interpolating polygonal approximation of it.



(a) Surface plot of y_h.

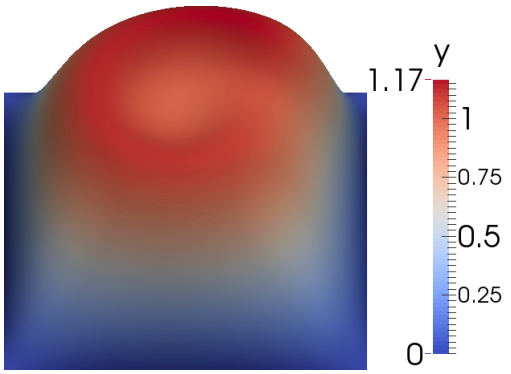
(b) Surface plot of $u_h = -\frac{1}{\nu}p_h$.



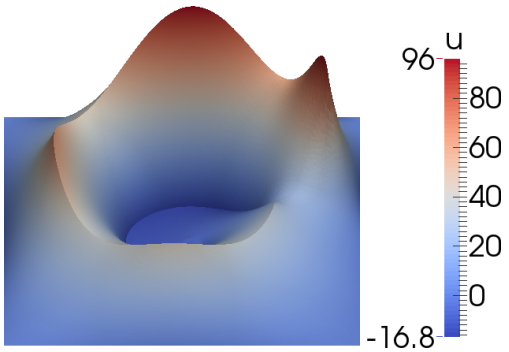
(c) Colour map of y_h.

(d) Colour map of $u_h = -\frac{1}{\nu}p_h$.

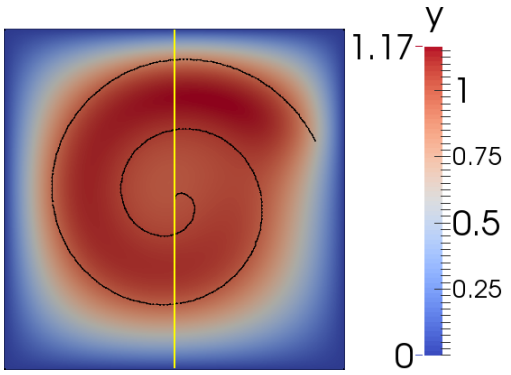
Figure 3.5: The solution to Example 3.23. We use Γ_σ, m_σ and $g_{\Gamma,\sigma}$ defined by Method 2. The black curve in Figure 3.5(c) is Γ and the dots and numerical values indicate the value of y_h at certain points on Γ. Not many points are included due to the symmetry of the solution.



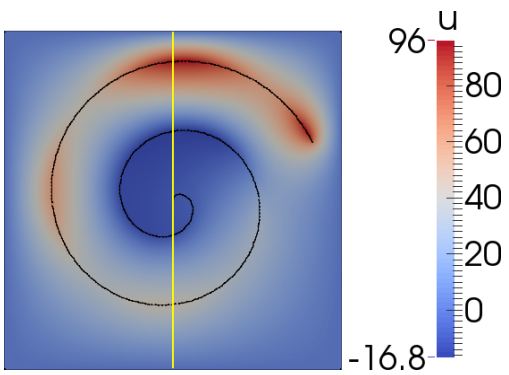
(a) Surface plot of y_h.



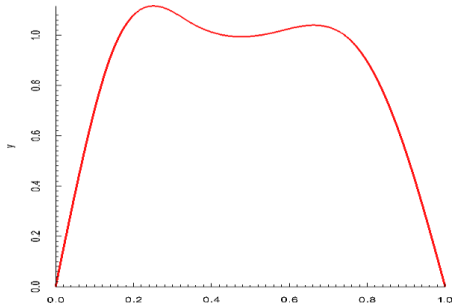
(b) Surface plot of $u_h = -\frac{1}{\nu}p_h$.



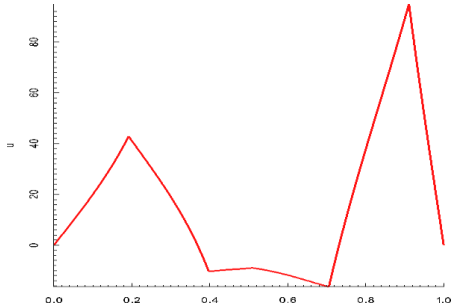
(c) Colour map of y_h.



(d) Colour map of $u_h = -\frac{1}{\nu}p_h$.

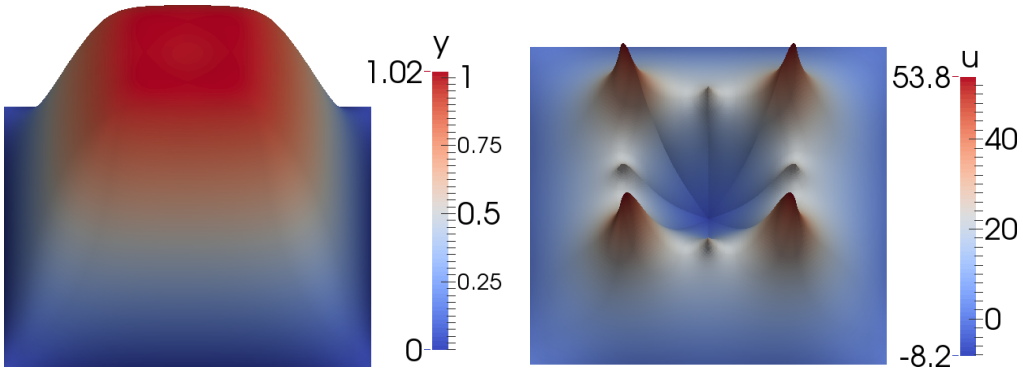


(e) Slice of y_h along the yellow line in Figure 3.6(c).



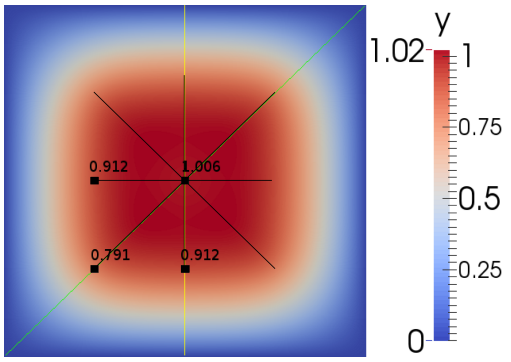
(f) Slice of $u_h = -\frac{1}{\nu}p_h$ along the yellow line in Figure 3.6(d).

Figure 3.6: The solution to Example 3.22. We use Γ_σ, m_σ and $g_{\Gamma,\sigma}$ defined by Method 2. The figure should be interpreted in the same way as Figure 3.5. In addition the yellow line indicates the slice used to produce Figures 3.6(e) and 3.6(f).

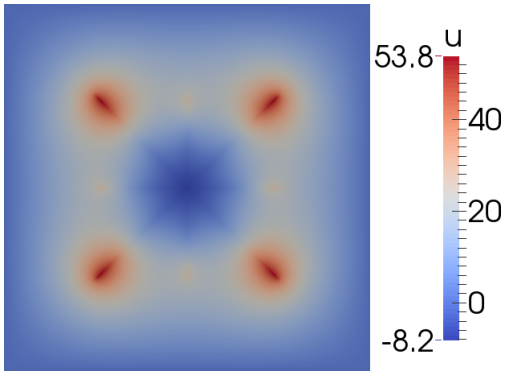


(a) Surface plot of y_h.

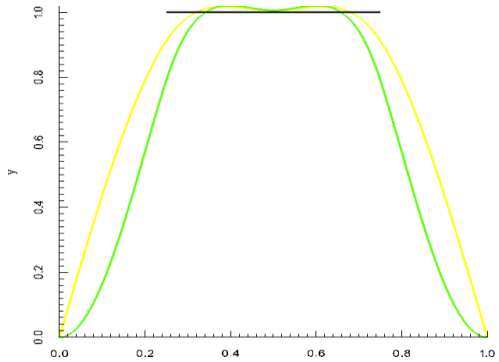
(b) Surface plot of $u_h = -\frac{1}{\nu}p_h$.



(c) Colour map of y_h.



(d) Colour map of $u_h = -\frac{1}{\nu}p_h$.



(e) Slices of $u_h = -\frac{1}{\nu}p_h$ along coloured lines in Figure 3.7(c). The green curve in Figure 3.7(e) has been scaled along the x-axis.

Figure 3.7: The solution to Example 3.24. We use Γ_σ, m_σ and $g_{\Gamma,\sigma}$ defined by Method 1. This Figure should be interpreted in the same way as Figure 3.6. Figure 3.7(e) shows some slices of y_h.

3.5.3 Comparison to optimal control at points

To finish this chapter we compare the solution of the line problem from Example 3.22 (shown in Figure 3.6) with the following point control problem.

Example 3.25. $\Omega = (0,1)^2$,

$$\Gamma = \{(0.5 + 0.327t \sin t, 0.5 + 0.327t \cos t) : t \in (0, 3.159)\},$$

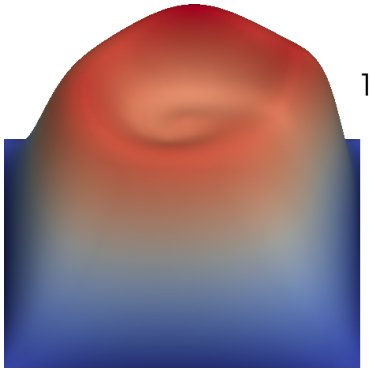
(i.e. a spiral), I is a set of 41 *evenly spaced points along* Γ, $g_\omega = 1$ *for all* $\omega \in I$, $\nu = 1e-4$, $b = -a = \infty$.

The theory for such problems is covered in the previous chapter. Note that we take the same parameter values as for the line problem except instead of a prescribed function $g_\Gamma = 1$, we have prescribed values of $g_\omega = 1$ at points along Γ. The solution of this problem can be seen in Figure 3.8.

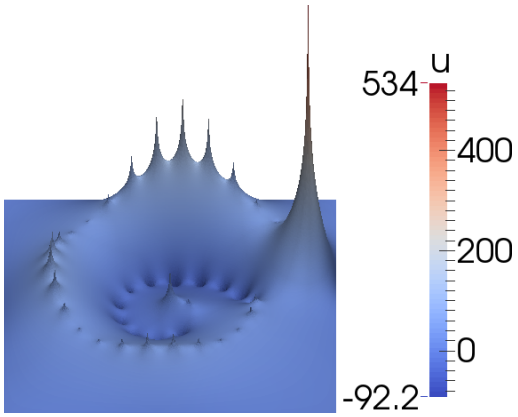
We see in Figure 3.8(e) that the point problem gets $y_h|_\Gamma$ closer to 1 than the line problem. However this is at the cost of $\|u_h\|_{L^2(\Omega)} = 36.5414$ for the point problem compared to $\|u_h\|_{L^2(\Omega)} = 28.0718$ for the line problem, and a what appears to be unbounded $\|u\|_\infty$.

Recall our observation from Section 3.4.2 that solutions of appropriately weighted discrete point control problems converge to the solution of a surface control problem. The points and weights we mentioned arose from Method 2. A simpler approach, which nevertheless works well in practice, is to choose an arbitrary triangulation of size h, then take $\lceil \frac{|\Gamma|}{h} \rceil$ (where $\lceil \cdot \rceil$ denotes the ceiling function) evenly spaced points along Γ and weight them by h. Given an arclength parameterisation of Γ, it is straightforward to adapt the implementation described in the previous chapter to do this.

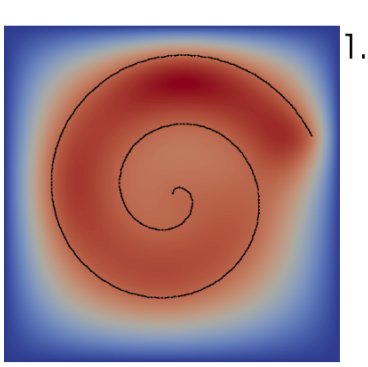
The solution to the point control problem resulting from this approach for $h = 0.00552$ can be seen in Figure 3.9. Comparing it to the solution of the line control problem using Method 2 (see Figure 3.6), we see that they are almost indistinguishable. A minor difference is that the ridge in u_h is slightly jagged, as the edges of the triangulation do not necessarily align with it, but as h is reduced this effect disappears.



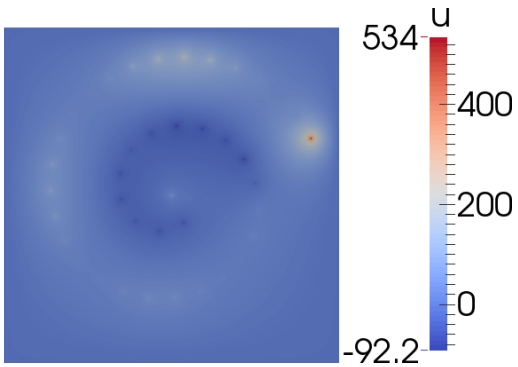
(a) Surface plot of y_h.



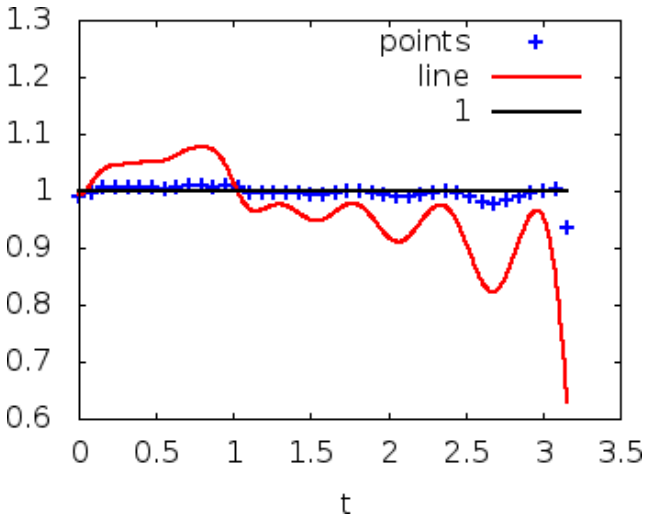
(b) Surface plot of $u_h = -\frac{1}{\nu}p_h$.



(c) Colour map of y_h. The black line is the curve Γ along which the points are evenly distributed.

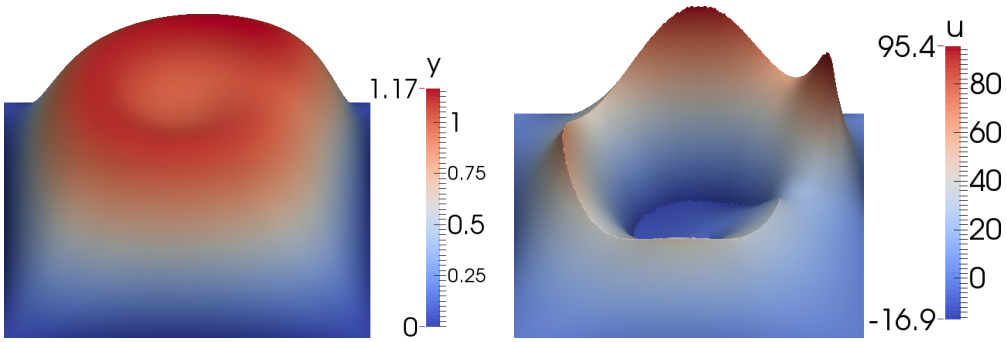


(d) Colour map of $u_h = -\frac{1}{\nu}p_h$.



(e) Comparison of y_h from the point control problem of Example 3.25 (shown above) evaluated on I, against y_h from the line control problem of Example 3.22 (shown in Figure 3.6) evaluated along Γ.

Figure 3.8: The solution of Example 3.25. Compare to Figure 3.6.



(a) Surface plot of y_h.

(b) Surface plot of $u_h = -\frac{1}{\nu}p_h$.

Figure 3.9: The solution to a point control that appears to closely approximate the solution to Example 3.22 (see Figure 3.6).

Chapter 4

Optimal control of elliptic variational inequalities at points

In this chapter we consider the optimal control of elliptic variational inequalities (VIs) with an objective functional containing the same point fidelity term as in Chapter 2 i.e. the distance between the state and prescribed values g_ω at a finite number of points $\omega \in I$. So for a bounded domain $\Omega \subset \mathbb{R}^2$ we consider a problem of the form

$$\text{Minimise } \frac{1}{2}\sum_{\omega\in I}(y(\omega) - g_\omega)^2 + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)}$$

subject to the state constraint

$$y \text{ solves an elliptic VI with forcing } \eta,$$

and the control constraint

$$a \leq \eta \leq b.$$

The notation here is the same as in Chapter 2, but we allow a and b to be functions with $a < b$ pointwise, and we have a variational inequality constraint instead of a PDE. We will state this problem precisely in Section 4.2.

Variational inequalities typically arise as equilibrium conditions for systems. For example, a standard example of an elliptic variational inequality models the resting position of an elastic membrane that is stretched over an obstacle. As a result, controlling such a variational inequality amounts to controlling the eventual position of the membrane. See [Hlaváček et al., 1985] for an example application. Our variant of the VI control problem has a point fidelity term, which has the

same motivation as controlling PDEs at points (considered in Chapter 2). More specifically, the problem is related to the inverse problem of parameter identification for a variational inequality with data at points, and a financial application of this is calibrating the volatility in the Black-Scholes equations for American options (see [Achdou, 2005]).

In contrast to the PDE control problems in previous chapters, the VI control problem is nonconvex and has poor differentiability properties, so it is harder. As a result we do not try to prove a priori error estimates (though in our numerical experiments in Section 4.7.1 we observe order h convergence). This numerical analysis is hard even with an $L^2(\Omega)$ fidelity term in the objective functional, and the theory is incomplete. For a start, the nonconvexity of the problem means that there is not a unique solution, which complicates the concept of an error estimate. Instead we concentrate on finding an efficient algorithm for solving the problem. In particular we consider an adaptive finite element approach.

The optimal control of VIs (with the standard objective functional containing an $L^2(\Omega)$-fidelity term, as in Chapter 1) has been well studied in both the finite dimensional case [Luo et al., 1996; Outrata et al., 1998] and the infinite dimensional case [Barbu and da Prato, 1984; Neittaanmäki et al., 2006]. In the literature this problem is commonly classified as a mathematical program with equilibrium constraints (MPEC), as variational inequalities often model an equilibrium condition. Unlike with the optimal control of PDEs in the previous chapters, poor differentiability properties of the control-to-state operator mean that standard methods cannot be used to derive necessary optimality conditions. To get around this, penalty methods (see e.g. [Barbu and da Prato, 1984]) and also generalised derivatives (see e.g. [Mignot, 1976]) have been applied to the problem. Also, alternative notions of stationarity have been derived in [Scheel and Scholtes, 2000] for finite dimensional problems, and in [Hintermüller and Kopacka, 2009; Outrata et al., 2011] in function spaces. The advantages of formulating stationarity conditions in function spaces are that this aids the design of mesh independent solution algorithms and it also allows us to derive a posteriori error estimators.

In particular, we can derive a dual-weighted goal-oriented a posteriori error estimator for $|J(y^*, u^*) - J(y_h^*, u_h^*)|$, the absolute difference in the objective functional evaluated at solutions to the continuous and discrete stationarity conditions. In our work the discrete stationarity conditions are derived by discretising with finite elements. This theory was developed and applied to the optimal control of PDEs in [Bangerth and Rannacher, 2003; Becker et al., 2000], then applied to the optimal control of variational inequalities (with the standard objective functional) in [Hin-

termüller et al., 2013]. There are alternative approaches to adaptive finite element methods (AFEMs) (see e.g. [Verfürth, 1996; Bangerth and Rannacher, 2003; Repin, 2008; Babuška et al., 2011]). In particular, dual-weighted residual error estimators for control of PDEs have been studied in [Hintermüller and Hoppe, 2008b, 2010a,b; Benedix and Vexler, 2009; Vexler and Wollner, 2008; Rösch and Wachsmuth, 2012; Günther and Hinze, 2008; Liu and Yan, 2001; Rösch and Wachsmuth, 2012], residual based estimators for control of PDEs have been studied in [Li et al., 2002; Hoppe and Kieweg, 2009, 2010; Hintermüller and Hoppe, 2008a; Hoppe et al., 2006; Liu and Yan, 2001], and residual based estimators for control of VIs have been studied in [Gaevskaya, 2013].

In the next section we will introduce some notation then in Section 4.2 we will state (often without proof) results on the analysis of our problem, which are taken from the joint work [Brett et al., 2013], and were proved mainly by Caroline Löbhard. We will see that the point evaluations in our objective functional cause mathematical difficulties, as we require the state space to embed into the space of continuous functions, which leads to reduced regularity of the adjoint variable. In Section 4.3, by considering a penalised and smoothed optimal control problem we introduce the concept of limiting ε-almost C-stationarity. We discretise this in Section 4.4 to get a discrete stationarity system, and then in Section 4.5 derive a dual-weighted goal-oriented estimator for the error in the objective functional. In Section 4.6 we use our analysis to define a solution algorithm for the discrete stationarity system, which is discretised using a finite element method. To finish, in Section 4.7 we show numerical results that demonstrate our adaptive finite element approach leads to an efficient method for finding controls that have small error in the objective functional. We also test an algorithm based on uniform refinement for finding discrete functions which closely approximate continuous optimal controls, and give a numerical example of control of a variational inequality on a line.

4.1 Notation

Let the domain Ω be a bounded open set in $\mathbb{R}^2$ with a Lipschitz boundary. In particular, we allow for domains with a smooth boundary and convex domains with polygonal boundaries, as we considered in previous chapters.

Let the differential operator A satisfy the same assumptions as in Section 2.1 but with $a_0 = 0$. We define the bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ corresponding

to A as before. We can also define the operator $\mathcal{A} : H_0^1(\Omega) \to H^{-1}(\Omega)$ that satisfies

$$\langle \mathcal{A}z, v\rangle_{H^{-1}(\Omega)} = a(z, v) \quad \forall z, v \in H_0^1(\Omega).$$

This operator agrees with A for suitably smooth functions so it is an extension of the domain of A that will simplify notation. For $\eta + f \in H^{-1}(\Omega)$ the following variational inequality has a unique solution (see e.g. Theorem 4:3.1 in [Rodrigues, 1987]):

$$\text{Find } y \in K \text{ such that } \langle \mathcal{A}y, v - y\rangle_{H^{-1}(\Omega)} \geq \langle \eta + f, v - y\rangle_{H^{-1}(\Omega)} \quad \forall v \in K. \tag{4.1}$$

Here $K := \{v \in H_0^1(\Omega) : v \geq 0 \text{ a.e. in } \Omega\}$ and throughout this chapter, for $v^* \in V^*$ and $v \in V$ we use $\langle v^*, v\rangle_{V^*} := v^*(v)$ to denote the duality pairing between a Banach space V and its dual space V^*. In contrast to Chapter 2, we include a forcing term in our exposition from the very beginning, as the map from the control u to the state y satisfying the variational inequality is already nonlinear. The variational inequality is equivalent to the following complementarity system with the slack variable ξ:

$$\begin{aligned} &\text{Find } y \in V \text{ such that } \xi = \mathcal{A}y - \eta - f \text{ in } H^{-1}(\Omega), \\ &y \geq 0 \text{ in } H_0^1(\Omega), \quad \xi \geq 0 \text{ in } H^{-1}(\Omega), \quad \langle \xi, y\rangle_{H^{-1}(\Omega)} = 0. \end{aligned} \tag{4.2}$$

By $\xi \geq 0$ in H^{-1} we mean $\langle \xi, v\rangle_{H^{-1}(\Omega)} \geq 0$ for all $v \in H_0^1(\Omega)$ with $v \geq 0$.

Under our assumption on the smoothness of the domain, if we have the additional regularity that $\eta + f \in L^2(\Omega)$ then there exists $Q > 2$ such that (4.1) satisfies $y \in W_0^{1,q}(\Omega)$ for all $q \in (2, Q)$ (see Proposition 2.1 in [Brett et al., 2013]). Rather than keeping the choice of q flexible as in Chapter 2, we will instead fix $q \in (2, Q)$. For the remainder of this chapter we will assume $f \in L^2(\Omega) \hookrightarrow W^{-1,q}(\Omega)$ and so we can denote by $S : L^2(\Omega) \to W_0^{1,q}(\Omega)$ the operator that maps $u \in L^2(\Omega)$ to the solution $y \in W_0^{1,q}(\Omega)$ of (4.1). This will be the control-to-state operator for our optimal control problem. Note that in comparison to Chapter 2, the range space of S is $W_0^{1,q}(\Omega)$. We do not continuously embed $W_0^{1,q}(\Omega)$ into $C_0(\Omega)$ and take this to be the state space, as $W_0^{1,q}(\Omega)$ turns out to be more convenient for the analysis.

4.2 Optimal control problem

With the above definitions we can now formulate the optimal control problem precisely as:

$$\min\ J(y,\eta) := \frac{1}{2}\sum_{\omega\in I}(y(\omega)-g_\omega)^2 + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)} \tag{4.3a}$$

$$\text{over } (y,\eta) \in W_0^{1,q}(\Omega)\times L^2(\Omega), \tag{4.3b}$$

$$\text{s.t. (4.1) holds} \tag{4.3c}$$

$$\text{and } \eta \in U_{ad} := \{L^2(\Omega) : a(x) \le \eta(x) \le b(x) \text{ a.e. } x \text{ in } \Omega\}. \tag{4.3d}$$

Here $f \in L^2(\Omega)$, I is a finite set of points in Ω, $g_\omega \in \mathbb{R}$ are prescribed values of the state at $\omega \in I$, and $\nu > 0$ represents the cost of the control. As we will not try to prove error estimates, it does not complicate notation to assume $a, b \in L^2(\Omega)$ with $a < b$ pointwise (in comparison to $a, b \in \mathbb{R}$ in previous chapters). We will also consider the case of $b = -a = \infty$ (i.e. $U_{ad} = L^2(\Omega)$).

We can use the control-to-state operator S that corresponds to (4.1) to define the reduced objective functional $\hat{J}(\eta) = J(S\eta, \eta)$, giving the equivalent optimisation problem

$$\min \hat{J}(\eta) = \frac{1}{2}\sum_{\omega\in I}(S\eta(\omega)-g_\omega)^2 + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)} \text{ over } \eta \in U_{ad}. \tag{4.4}$$

Lemma 4.1. *Problem (4.4) has a solution $u \in U_{ad}$ and so (4.3) has a solution $(Su, u) \in W_0^{1,q}(\Omega)\times U_{ad}$).*

Proof. Note that J is continuous and convex, and hence weakly lower semicontinuous. If a sequence $\eta_k \in L^2(\Omega)$ converges weakly to η in $L^2(\Omega)$, then $S\eta_k$ has a subsequence converging weakly to $S\eta$ in $W_0^{1,q}(\Omega)$ (see Lemma 2.4 in [Hintermüller et al., 2013]). So the standard argument (see e.g. [Tröltzsch, 2010]) gives existence of a solution to (4.4). □

Despite existence of a solution, this problem is hard to solve in practice; $\hat{J}$ is nonconvex and nondifferentiable because S is nonlinear and nondifferentiable. In particular, the nondifferentiability means standard techniques cannot be used to derive first order stationarity conditions for the problem. This motivates us to approximate it in the next section with a problem which is differentiable and to which standard theory applies.

4.3 Penalised optimal control problem with smoothed objective functional

In this section we introduce a penalised optimal control problem with a smoothed objective functional for which standard theory can be used to derive stationarity conditions. In order to do this we replace the VI constraint with an approximating PDE constraint and a modified objective functional. We justify this by proving that solutions of the approximating optimal control problem converge to solutions of the original optimal control problem (4.3). We will sometimes refer the reader to [Brett et al., 2013] for more details.

The variational inequality (4.1) can be approximated by the following weak formulation of a semilinear PDE when the parameter γ is large:

$$\text{Find } y \in H_0^1(\Omega) \text{ s.t. } \mathcal{A}y - \gamma \max(0, -y) = \eta + f \text{ in } H^{-1}(\Omega). \tag{4.5}$$

This has a unique solution as it is the first order optimality condition of a strictly convex optimisation problem, and it approximates the variational inequality in some sense (see [Brett et al., 2013]). However the solution operator of (4.5) is still not differentiable due to the kink in $\max(0, \cdot)$. We can fix this by replacing $\max(0, \cdot)$ with a function $\max_\varepsilon(0, \cdot)$ that smooths the kink and has the property that

$$\max_\varepsilon(0, \cdot) \to \max(0, \cdot) \text{ a.e. as } \varepsilon \to 0.$$

The smoothing that we use is the C^1 local smoothing

$$\max_\varepsilon^l(0, t) := \begin{cases} 0 & \text{if } t \leq -\varepsilon, \\ \frac{t^2}{4\varepsilon} + \frac{t}{2} + \frac{\varepsilon}{4} & \text{if } t \in (-\varepsilon, \varepsilon), \\ t & \text{if } t \geq \varepsilon. \end{cases}$$

(see e.g. [Hintermüller and Kopacka, 2011]).

With this we approximate the VI (4.1) by the following weak formulation of a semilinear PDE:

$$\text{Find } y \in H_0^1(\Omega) \text{ s.t. } \mathcal{A}y - \gamma \max_\varepsilon(0, -y) = \eta + f \text{ in } H^{-1}(\Omega). \tag{4.6}$$

This has a unique solution as it can also be viewed as the first order optimality condition of a strictly convex optimisation problem. Moreover for $\eta + f \in L^2(\Omega)$ we have the additional regularity that $y \in W_0^{1,q}(\Omega)$. To simplify the parameter space

we fix ε to be a function of γ with the following property.

Assumption 4.2. *For $\gamma > 0$, let $\varepsilon(\gamma)$ be such that $\lim_{\gamma\to\infty} \gamma\varepsilon(\gamma) = 0$.*

This assumption is sufficient for our analysis. We can now denote by $S^\gamma : L^2(\Omega) \to W_0^{1,q}(\Omega)$ the operator that maps $\eta \in L^2(\Omega)$ to the solution $y \in W_0^{1,q}(\Omega)$ of (4.6) for a given γ and $\varepsilon = \varepsilon(\gamma)$ satisfying Assumption 4.2. We can then show that

$$\|S^\gamma \eta\|_{W_0^{1,q}(\Omega)} \leq C\|\eta + f\|_{L^2(\Omega)}$$

(see Proposition 2.6 in [Brett et al., 2013]) and also that the PDE (4.6) approximates the VI (4.1) in the following sense.

Theorem 4.3. *Let $(\eta_k)_{k\in\mathbb{N}}$ be a sequence in U_{ad} converging weakly in $L^2(\Omega)$ to $\eta \in U_{ad}$ and let $\gamma_k \to \infty$. Then there exists a (relabelled) subsequence of $(\eta_k)_{k\in\mathbb{N}}$ and $(\gamma_k)_{k\in\mathbb{N}}$ such that for $k \to \infty$,*

$$S^{\gamma_k}\eta_k \to S\eta \text{ in } W_0^{1,q}(\Omega).$$

Note that Assumption 4.2 implicitly holds due to how S^γ is defined.

Proof. See Theorem 2.8 in [Brett et al., 2013]. □

Note that S^γ is differentiable, so we could use standard theory to derive stationary conditions for the optimal control of this operator. We also smooth the objective functional J as this turns out to be necessary to prove convergence to the original optimal control problem. So we consider the following penalised optimal control problem with smoothed objective functional:

$$\min\ J_r(y,\eta) = \frac{1}{2\,|B_r(0)|}\sum_{\omega\in I}\|y - g_\omega\|^2_{L^2(B_r(\omega))} + \frac{\nu}{2}\|\eta\|^2_{L^2(\Omega)} \tag{4.7a}$$

$$\text{over } (y,\eta) \in W_0^{1,q}(\Omega) \times L^2(\Omega), \tag{4.7b}$$

$$\text{s.t. (4.6) holds} \tag{4.7c}$$

$$\text{and } \eta \in U_{ad}, \tag{4.7d}$$

where $r > 0$ and $B_r(\omega) := \{x \in \Omega : |x - \omega| < r\}$. The smoothing in the objective is justified by the fact that for $y \in W_0^{1,q}(\Omega)$,

$$\frac{1}{|B_r(\omega)|}\|y - g_\omega\|^2_{L^2(B_r(\omega))} \to (y(\omega) - g_\omega)^2 \text{ as } r \to 0.$$

This problem has a solution (see e.g. Theorem 1.45 in [Hinze et al., 2009]). The penalised optimal control problem with smoothed objective functional (4.7) approximates the original optimal control problem (4.3) in the following sense.

Theorem 4.4. *Let $(\gamma_k)_{k\in\mathbb{N}} \subset \mathbb{R}^{>0}$ tend to infinity, $(r_k)_{k\in\mathbb{N}} \subset \mathbb{R}^{>0}$ converge to zero, and $\varepsilon(\gamma_k)$ satisfy Assumption 4.2. Denote solutions of (4.7) with $(\gamma, \varepsilon, r) = (\gamma_k, \varepsilon(\gamma_k), r_k)$ by $(y_k, u_k) = (S^{\gamma_k} u_k, u_k)$. Then there exists a (relabelled) subsequence (y_k, u_k) and a minimiser (y, u) of (4.3) such that*

$$y_k \to y \quad \text{in } W_0^{1,q}(\Omega), \quad u_k \rightharpoonup u \in L^2(\Omega).$$

4.3.1 First order stationarity conditions

The penalised optimal control problem (4.7) is still nonconvex, so in order to solve it in practice we derive some stationarity conditions which are suitable for numerical solution. We then define a notion of stationarity for the original optimal control problem (4.3) that is motivated by the literature. To finish we prove convergence of stationary points of the penalised problem with smoothed objective functional to stationary points of the original problem.

For the penalised optimal control problem with smoothed objective functional

The control-to-state operator S^γ of (4.7) is differentiable so we can derive first order stationarity conditions using standard theory (see [Brett et al., 2013]): If $(y_k, u_k) \in H_0^1(\Omega) \times L^2(\Omega)$ is a minimiser of (4.7) with parameters $(\gamma, \varepsilon, r) = (\gamma_k, \varepsilon(\gamma_k), r_k)$ then there exists a $p_k \in H_0^1(\Omega)$ such that

$$\mathcal{A}^* p_k + \gamma_k \max'_{\varepsilon(\gamma_k)}(0, -y_k) p_k = \frac{1}{|B_{r_k}(0)|} \sum_{\omega \in I} (y_k - g_\omega) \chi_{B_{r_k}}(\omega) \quad \text{in } H^{-1}(\Omega), \tag{4.8a}$$

$$u_k \in U_{ad}, \quad (p_k + \nu u_k, v - u_k)_{L^2(\Omega)} \geq 0 \quad \forall v \in U_{ad}. \tag{4.8b}$$

Here $\mathcal{A}^* : H^{-1} \to H_0^1(\Omega)$ is the adjoint of $\mathcal{A}$ defined by $\langle v^*, \mathcal{A}^* w \rangle_{H^{-1}(\Omega)} = \langle \mathcal{A} v^*, w \rangle_{H^{-1}(\Omega)}$ for all $v^* \in H^{-1}(\Omega), w \in H_0^1(\Omega)$ and χ_B denotes an indicator function for a set B. We say (y_k, u_k, p_k) is a stationary point for (4.7) with parameters $(\gamma, \varepsilon, r) = (\gamma_k, \varepsilon(\gamma_k), r_k)$ if $y_k = S^{\gamma_k} u_k$ and (4.8) holds.

Remark 4.5. *Note that (4.8b) can equivalently be written in the following three*

ways:

$$u_k = -\frac{1}{\nu}p_k + \max(0, a + \frac{1}{\nu}p_k) - \max(0, -\frac{1}{\nu}p_k - b) \quad \text{a.e. in } \Omega,$$

$$u_k = \prod\nolimits_{U_{ad}} \left(-\frac{1}{\nu}p_k\right),$$

$$\exists \sigma_{a,k}, \sigma_{b,k} \in L^2(\Omega) \text{ s.t.}$$

$u_k \in U_{ad}$, $\sigma_{a,k} - \sigma_{b,k} = p_k + \nu u_k$, $\sigma_{a,k} \geq 0$, $\sigma_{b,k} \geq 0$, $\sigma_{a,k}(a - u_k) = \sigma_{b,k}(b - u_k) = 0$.

Here $\prod_{U_{ad}}$ *denotes the* $L^2(\Omega)$*-projection onto* U_{ad}*.*

For the optimal control problem

We can use the ideas in [Hintermüller and Kopacka, 2009] to define a notion of stationarity for the original optimal control problem (4.3) such that stationary points of the penalised optimal control problem with smooth objective functional (4.7) converge to stationary points of (4.3).

Definition 4.6. *We define* $(y, u) \in W_0^{1,q}(\Omega) \times L^2(\Omega)$ *to be limiting* ε*-almost C-stationary for* (4.3) *with slack variable* $\xi = \mathcal{A}y - u - f \in W^{-1,q}(\Omega)$ *if there exist multipliers* $(p, \lambda) \in W_0^{1,q'}(\Omega) \times W^{-1,q'}(\Omega)$ *such that*

$$(4.1) \text{ holds,} \tag{4.9a}$$

$$\mathcal{A}^* p - \lambda - \sum_{\omega \in I} (y(\omega) - g_\omega)\delta_\omega = 0 \text{ in } W^{-1,q'}(\Omega), \tag{4.9b}$$

$$u \in U_{ad}, \quad (p + \nu u, v - u)_{L^2(\Omega)} \geq 0 \quad \forall v \in U_{ad}, \tag{4.9c}$$

$$\langle \lambda, y \rangle_{W^{-1,q'}(\Omega)} = 0, \tag{4.9d}$$

$$\forall \tau > 0 \quad \exists E_\tau \subset \Omega^+ := \{y > 0\} \text{ such that } \left|\Omega^+ \setminus E_\tau\right| < \tau \text{ and}$$

$$\forall \varphi \in L^\infty(\Omega) \text{ with } \varphi|_{\Omega \setminus E_\tau} = 0, \quad \langle \lambda, \varphi \rangle_{L^\infty(\Omega)^*} = 0, \tag{4.9e}$$

$$\exists \lambda_k \rightharpoonup^* \lambda \text{ in } L^\infty(\Omega)^*, p_k \rightharpoonup p \text{ in } W_0^{1,q'}(\Omega) \text{ s.t.}$$

$$\limsup_{k \to \infty} \langle \lambda_k, p_k \rangle_{H^{-1}(\Omega)} \geq 0, \tag{4.9f}$$

$$\langle \xi, p \rangle_{W^{-1,q}(\Omega)} = 0. \tag{4.9g}$$

Here δ_ω denotes a delta function centred at ω, which belongs to $W^{-1,q'}(\Omega)$ as $\langle \delta_\omega, v \rangle_{W^{-1,q'}(\Omega)} = \int_\Omega v \delta_\omega \mathrm{d}x = v(\omega)$ for $v \in W_0^{1,q}(\Omega)$.

Theorem 4.7. *Let* $(\gamma_k)_{k \in \mathbb{N}} \subset \mathbb{R}^{>0}$ *tend to infinity,* $(r_k)_{k \in \mathbb{N}} \subset \mathbb{R}^{>0}$ *converge to zero, and* $\varepsilon(\gamma_k)$ *satisfy Assumption 4.2. Let* (y_k, u_k) *be stationary for (4.7) with multiplier* p_k *for* $(\gamma, \varepsilon, r) = (\gamma_k, \varepsilon(\gamma_k), r_k)$ *and assume that* $(\|u_k\|_{L^2(\Omega)})_{k \in \mathbb{N}}$ *is bounded. Then*

there exists a limiting ε-almost C-stationary point $(y,u) \in W_0^{1,q}(\Omega) \times H_0^1(\Omega)$ for problem (4.3) with slack variable $\xi \in L^2(\Omega)$ and multipliers $(p,\lambda) \in H_0^1(\Omega) \times L^\infty(\Omega)^*$ and a (relabelled) subsequence such that

$$\begin{aligned}
&y_k \to y && \text{in } W_0^{1,q}(\Omega),\\
&u_k \rightharpoonup u && \text{in } H_0^1(\Omega),\\
&\xi_k := \gamma_k \max_{\varepsilon_k}(0,-y_k) \rightharpoonup \xi && \text{in } L^2(\Omega),\\
&p_k \rightharpoonup p && \text{in } H_0^1(\Omega),\\
&\lambda_k := -\gamma_k \max'_{\varepsilon_k}(0,-y_k)p_k \rightharpoonup^* \lambda && \text{in } L^\infty(\Omega)^*.
\end{aligned}$$

4.4 Discretisation

In this section we formulate discrete stationarity conditions for the penalised optimal control problem with smoothed objective functional (4.7) and the original optimal control problem (4.3).

Let $V_h \subset W_0^{1,q}(\Omega)$ be a finite dimensional space. We intend this to be a finite element space but do not assume this quite yet. As in the previous chapters we use the variational discretisation concept from [Hinze, 2005] and so we have an implicitly discretised control that belongs to $U_{ad} \subset L^2(\Omega)$.

We define discrete stationary points of (4.7) as $(y_h, u_h, p_h) \in V_h \times L^2(\Omega) \times V_h$ such that

$$a(y_h, v_h) + (-\gamma \max_{\varepsilon(\gamma)}(0,-y_h) - u_h - f, v_h)_{L^2(\Omega)} = 0 \tag{4.10a}$$

$$a(v_h, p_h) + \gamma(\max'_{\varepsilon(\gamma)}(0,-y_h)p_h, v_h)_{L^2(\Omega)} - \sum_{\omega\in I}(y_h(\omega) - g_\omega)v_h(\omega) = 0 \tag{4.10b}$$

$$u_h \in U_{ad}, \quad (p_h + \nu u_h, v - u_h)_{L^2(\Omega)} \geq 0 \tag{4.10c}$$

for all $v_h \in V_h$ and $v \in U_{ad}$. Note that this discretisation contains the term

$$\sum_{\omega\in I}(y_h(\omega) - g_\omega)v_h(\omega). \tag{4.11}$$

There are other possible options here, such as

$$\Big(\frac{1}{|B_r(0)|}\sum_{\omega\in I}(y_h - g_\omega)\chi_{B_r(\omega)}, v_h\Big)_{L^2(\Omega)}. \tag{4.12}$$

For many choices of discrete space, such as finite element spaces, it is difficult to compute (4.12). To get around this for finite element spaces, instead of integrating

over a ball we could integrate over a polygon of simplices in the triangulation (which we will introduce at the end of this section). For example, set $B_{\omega,h} := \bigcup_{\omega \in T} T$, where T are simplices in the triangulation, and calculate

$$\Big(\frac{1}{|B_{\omega,h}|} \sum_{\omega \in I} (y_h - g_\omega) \chi_{B_{\omega,h}}, v_h \Big)_{L^2(\Omega)}.$$

This quantity is easier to compute than (4.12), however we are only able to send $r \to 0$ at the same rate as h. If we want to send $r \to 0$ slower than h then we need the polygons to contain more and more simplices, which requires a complicated implementation. Also with this method it is necessary to ensure all evaluation points are vertices of the triangulation, otherwise slight perturbations of the evaluation points could lead to quite different numerical solutions.

If we want to send $r \to 0$ faster than h then in the limit we obtain (4.11). This is the discretisation we actually use. By computing this quantity for all h we remove some smoothing which was necessary to prove convergence in the analysis, however this does not seem to cause a problem numerically. It is also implementationally straightforward in finite element spaces.

We define discrete stationary points of (4.3) in the following way.

Definition 4.8. *$(y_h, u_h) \in V_h \times L^2(\Omega)$ is a discrete stationary point for (4.3) with slack variable $\xi_h \in V_h^*$ defined by*

$$\langle \xi_h, v_h \rangle_{V_h^*} = a(y_h, v_h) - (u_h + f, v_h)_{L^2(\Omega)} \quad \forall v_h \in V_h$$

if there exist multipliers $(p_h, \lambda_h) \in V_h \times V_h^*$ such that

$$y_h \geq 0, \quad a(y_h, v_h - y_h) - (u_h + f, v_h - y_h)_{L^2(\Omega)} \geq 0 \qquad \forall v_h \in V_h, \tag{4.13a}$$

$$a(v_h, p_h) - \langle \lambda_h, v_h \rangle_{V_h^*} - \sum_{\omega \in I} (y_h(\omega) - g_\omega) v_h(\omega) = 0 \qquad \forall v_h \in V_h, \tag{4.13b}$$

$$u_h \in U_{ad}, \quad (p_h + \nu u_h, v - u_h) \geq 0 \qquad \forall v \in U_{ad}, \tag{4.13c}$$

$$\langle \lambda_h, y_h \rangle_{V_h^*} = 0, \tag{4.13d}$$

$$\forall v_h \in V_h \text{ with } v_h|_{\{y_h = 0\}} = 0, \quad \langle \lambda_h, v_h \rangle_{V_h^*} = 0, \tag{4.13e}$$

$$\langle \xi_h, p_h \rangle_{V_h^*} = 0, \tag{4.13f}$$

$$\langle \lambda_h, p_h \rangle_{V_h^*} \geq 0. \tag{4.13g}$$

Each line of (4.13) is a discrete version of the corresponding line in (4.9). We will now write an equivalent definition which is better suited for deriving the primal-dual weighted error estimator in the next section.

Definition 4.9. *$(y_h, u_h) \in V_h \times L^2(\Omega)$ is a discrete stationary point for (4.3) with slack variable $\xi_h \in V_h^*$ defined by*

$$\langle \xi_h, v_h \rangle_{V_h^*} = a(y_h, v_h) - (u_h + f, v_h)_{L^2(\Omega)} \quad \forall v_h \in V_h \tag{4.14}$$

if there exist multipliers $p_h, \mu_h, \sigma_{a,h}, \sigma_{b,h} \in V_h$ and $\lambda_h \in V_h^*$ such that

$$\xi_h \geq 0,\ y_h \geq 0,\ \langle \xi_h, y_h \rangle_{V_h^*} = 0, \tag{4.15a}$$

$$a(v_h, p_h) - \langle \lambda_h, v_h \rangle_{V_h^*} - \sum_{\omega \in I} (y_h(\omega) - g_\omega) v_h(\omega) = 0 \qquad \forall v_h \in V_h, \tag{4.15b}$$

$$(p_h + \nu u_h - \sigma_{a,h} + \sigma_{b,h}, v)_{L^2(\Omega)} = 0 \qquad \forall v \in L^2(\Omega), \tag{4.15c}$$

$$(\mu_h - p_h, v_h)_{L^2(\Omega)} = 0 \qquad \forall v_h \in V_h \tag{4.15d}$$

$$a \leq u_h,\ \sigma_{a,h} \geq 0,\ (a - u_h, \sigma_{a,h})_{L^2(\Omega)} = 0, \tag{4.15e}$$

$$u_h \leq b,\ \sigma_{b,h} \geq 0,\ (u_h - b, \sigma_{b,h})_{L^2(\Omega)} = 0, \tag{4.15f}$$

$$\langle \lambda_h, y_h \rangle_{V_h^*} = 0, \tag{4.15g}$$

$$\forall v_h \in V_h \text{ with } v_h|_{\{y_h = 0\}} = 0, \quad \langle \lambda_h, v_h \rangle_{V_h^*} = 0, \tag{4.15h}$$

$$\langle \xi_h, \mu_h \rangle_{V_h^*} = 0, \tag{4.15i}$$

$$\langle \lambda_h, \mu_h \rangle_{V_h^*} \geq 0. \tag{4.15j}$$

Here we have rewritten (4.13a) using the slack variable ξ_h, used Remark 4.5 to replace (4.13c), and introduced the variable $\mu_h = p_h$. This will allow us to relate the discrete system to the MPCC-Lagrangian in the next section.

4.4.1 Finite element discretisation

As in previous chapters we will further assume that Ω is convex and take V_h to be a space of piecewise linear globally continuous finite elements which vanish on the boundary. In particular, take a polyhedral approximation Ω_h of Ω with a conforming shape-regular triangulation T_h, then define

$$V_h := \{v_h \in C_0(\Omega) : v|_T \in P_1(T) \text{ for all } T \in T_h \text{ and } v_h|_{\Omega \setminus \Omega_h} = 0\},$$

where $P_1(T)$ is the set of affine functions over T. The convexity ensures that $V_h \subset W_0^{1,q}(\Omega)$. See Section 2.3 for more details.

4.5 Primal-dual weighted error estimator

In order for discrete stationary points to closely approximate limiting ε-almost C-stationary points, we expect that h needs to be small. This corresponds to many elements in our discretisation, which leads to a high computational cost. Our goal is to find points (y, u) minimising J, therefore we design an estimator that indicates whether there is a high local error in the quantity $|J(y, u) - J(y_h, u_h)|$, where (y_h, u_h) is a discrete stationary point. We can then concentrate our grid refinement on these areas of high error, hence reducing unnecessary computational cost. The following calculations mimic those in [Hintermüller et al., 2013] and [Hintermüller and Hoppe, 2008b], which are based on the dual-weighted residual-based error approach of Bangerth and Rannacher (see for example [Bangerth and Rannacher, 2003]).

4.5.1 Abstract error representation

To begin with we derive an error representation for the discretisation with abstract discrete spaces from Section 4.4.

Let $x = (y, u, \xi, p) \in W_0^{1,q}(\Omega) \times L^2(\Omega) \times W^{-1,q}(\Omega) \times W_0^{1,q'}(\Omega)$ and $m = (\lambda, \mu, \sigma_a, \sigma_b) \in W^{-1,q'}(\Omega) \times W_0^{1,q'}(\Omega) \times L^2(\Omega) \times L^2(\Omega)$ and define the MPCC-Lagrangian as

$$\begin{aligned}\mathcal{L}(x, m) := J(y, u) + a(y, p) - \langle \xi, p\rangle_{W^{-1,q}(\Omega)} - (u + f, p)_{L^2(\Omega)} \\ - \langle \xi, \mu\rangle_{W^{-1,q}(\Omega)} - \langle \lambda, y\rangle_{W^{-1,q'}(\Omega)} \\ - (u - a, \sigma_a)_{L^2(\Omega)} - (b - u, \sigma_b)_{L^2(\Omega)}.\end{aligned}$$

Note how p acts as a Lagrange multiplier for the equality constraint $\mathcal{A}y - \xi - u - f = 0$, and $\mu, \lambda, \sigma_a, \sigma_b$ act as Lagrange multipliers for the inequality constraints $\xi \geq 0, y \geq 0, u - a \geq 0, b - u \geq 0$ respectively.

From now on let $x^* = (y^*, u^*, \xi^*, p^*)$ and $m^* = (\lambda^*, \mu^*, \sigma_a^*, \sigma_b^*)$ denote a limiting ε-almost C-stationary point along with the associated slack variable and multipliers. Similarly let x_h^* and m_h^* denote the variables for a discrete stationary point. We seek a representation of $J(y_h^*, u_h^*) - J(y^*, u^*)$.

Taylor expanding $\mathcal{L}(x_h^*, m_h^*)$ at x^* gives

$$\mathcal{L}(x_h^*, m_h^*) = \mathcal{L}(x^*, m_h^*) + \nabla_x \mathcal{L}(x^*, m_h^*)(x_h^* - x^*) + \frac{1}{2}\nabla_{xx}\mathcal{L}(x_h^* - x^*, x_h^* - x^*). \quad (4.16)$$

$\mathcal{L}(x, m)$ is a quadratic functional in x so higher order derivatives are zero. As its sec-

ond Fréchet derivative is independent of (x, m) we have abbreviated $\nabla_{xx}\mathcal{L}(x, m)(\delta x, \delta m)$ by $\nabla_{xx}\mathcal{L}(\delta x, \delta m)$. Note that $\mathcal{L}(x_h^*, m_h^*) = J(y_h^*, u_h^*)$ and

$$\begin{aligned}\mathcal{L}(x^*, m_h^*) =& J(y^*, u^*) - \langle \lambda_h^*, y^* \rangle_{W^{-1,q'}(\Omega)} - \langle \xi^*, \mu_h^* \rangle_{W^{-1,q}(\Omega)} \\ &- (u^* - a, \sigma_{a,h}^*)_{L^2(\Omega)} - (b - u^*, \sigma_{b,h}^*)_{L^2(\Omega)}, \end{aligned} \tag{4.17}$$

so (4.16) can be used to get an expression for $J(y_h^*, u_h^*) - J(y^*, u^*)$.

We need a representation for the Hessian $\frac{1}{2}\nabla_{xx}\mathcal{L}(x_h^* - x^*, x_h^* - x^*)$ in (4.16). Taylor expanding $\nabla_x \mathcal{L}(x_h^*, m_h^*)(x_h^* - x^*)$ at x^* gives

$$\nabla_x \mathcal{L}(x_h^*, m_h^*)(x_h^* - x^*) = \nabla_x \mathcal{L}(x^*, m_h^*)(x_h^* - x^*) + \nabla_{xx}\mathcal{L}(x_h^* - x^*, x_h^* - x^*).$$

Rearranging to get a representation of the Hessian and substituting this into (4.16) we get

$$\mathcal{L}(x_h^*, m_h^*) = \mathcal{L}(x^*, m_h^*) + \frac{1}{2}\nabla_x \mathcal{L}(x^*, m_h^*)(x_h^* - x^*) + \frac{1}{2}\nabla_x \mathcal{L}(x_h^*, m_h^*)(x_h^* - x^*). \tag{4.18}$$

We can calculate $\nabla_x \mathcal{L}(x_h^*, m_h^*)(x_h^* - x^*)$ and $\nabla_x \mathcal{L}(x^*, m_h^*)(x_h^* - x^*)$:

$$\begin{aligned}&\nabla_x \mathcal{L}(x^*, m_h^*)(x_h^* - x^*) \\ &\quad = \langle \lambda^*, y_h^* \rangle_{W^{-1,q'}(\Omega)} + \langle \lambda_h^*, y^* \rangle_{W^{-1,q'}(\Omega)} + \langle \xi^*, \mu_h^* \rangle_{W^{-1,q}(\Omega)} + \langle \xi_h^*, \mu^* \rangle_{W^{-1,q}(\Omega)} \\ &\quad + (u_h^* - a, \sigma_a^*)_{L^2(\Omega)} + (u^* - a, \sigma_{a,h}^*)_{L^2(\Omega)} \\ &\quad + (b - u_h^*, \sigma_b^*)_{L^2(\Omega)} + (b - u^*, \sigma_{b,h}^*)_{L^2(\Omega)},\end{aligned}$$

and

$$\begin{aligned}&\nabla_x \mathcal{L}(x_h^*, m_h^*)(x_h^* - x^*) \\ &\quad = a(y_h^*, p_h^* - p^*) - \langle \xi_h^*, p_h^* - p^* \rangle_{W^{-1,q}(\Omega)} - (u_h^* + f, p_h^* - p^*)_{L^2(\Omega)} \\ &\quad + a(y_h^* - y^*, p_h^*) - \langle \lambda_h^*, y_h^* - y^* \rangle_{W^{-1,q'}(\Omega)} - \sum_{\omega \in I} (y_h^*(\omega) - g_\omega)(y_h^*(\omega) - y^*(\omega)) \\ &\quad + (-p_h^* + \nu u_h^* - \sigma_{a,h}^* + \sigma_{b,h}^*, u_h^* - u^*)_{L^2(\Omega)} \\ &\quad + \langle \xi_h^* - \xi^*, p_h^* - \mu_h^* \rangle_{W^{-1,q}(\Omega)}.\end{aligned}$$

Note that due to the discrete stationarity system, in fact

$$\nabla_x \mathcal{L}(x_h^*, m_h^*)(x_h^* - x^*) = \nabla_x \mathcal{L}(x_h^*, m_h^*)(x_h^* - \delta x_h)$$

for all $\delta x_h = (\delta y_h, \delta u_h, \delta \xi_h, \delta p_h) \in V_h \times L^2(\Omega) \times V_h^* \times V_h$. So substituting these calculations along with (4.17) into (4.18) gives the following result.

Theorem 4.10. *If (y^*, u^*) is a limiting ε-almost C-stationary point with slack variable ξ^* and multipliers $p^*, \lambda^*, \mu^*, \sigma_a^*, \sigma_b^*$, and (y_h^*, u_h^*) is a discrete stationary point with slack variable ξ_h^* and multipliers $p_h^*, \lambda_h^*, \mu_h^*, \sigma_{a,h}^*, \sigma_{b,h}^*$, then*

$$
\begin{aligned}
&2(J(y^*, u^*) - J(y_h^*, u_h^*)) \\
&= a(y_h^*, p^* - \delta p_h) - \langle \xi_h^*, p^* - \delta p_h \rangle_{W^{-1,q'}(\Omega)} - (u_h^* + f, p^* - \delta p_h^*)_{L^2(\Omega)} && \text{(4.19a)} \\
&+ a(y^* - \delta y_h, p_h^*) - \langle \lambda_h^*, y^* - \delta y_h \rangle_{W^{-1,q'}(\Omega)} - \sum_{\omega \in I} (y_h^*(\omega) - g_\omega)(y^*(\omega) - \delta y_h(\omega)) \\
& && \text{(4.19b)} \\
&+ (p_h^* + \nu u_h^* - \sigma_{a,h}^* + \sigma_{b,h}^*, u^* - \delta u_h)_{L^2(\Omega)} && \text{(4.19c)} \\
&+ \langle \xi^* - \delta \xi_h, p_h^* - \mu_h^* \rangle_{W^{-1,q}(\Omega)} && \text{(4.19d)} \\
&+ \langle \lambda_h^*, y^* \rangle_{W^{-1,q'}(\Omega)} + \langle \xi_h^*, \mu^* \rangle_{W^{-1,q}(\Omega)} && \text{(4.19e)} \\
&+ \langle \lambda^*, y_h^* \rangle_{W^{-1,q'}(\Omega)} + \langle \xi^*, \mu_h^* \rangle_{W^{-1,q}(\Omega)} && \text{(4.19f)} \\
&+ (u^* - a, \sigma_{a,h}^*)_{L^2(\Omega)} - (u_h^* - a, \sigma_a^*)_{L^2(\Omega)} && \text{(4.19g)} \\
&+ (b - u^*, \sigma_{b,h}^*)_{L^2(\Omega)} - (b - u_h^*, \sigma_b^*)_{L^2(\Omega)}, && \text{(4.19h)}
\end{aligned}
$$

for all $\delta x_h = (\delta y_h, \delta u_h, \delta \xi_h, \delta p_h) \in V_h \times L^2(\Omega) \times V_h^* \times V_h$.

Note that the quantity on the right hand side of (4.19) cannot be computed numerically, as limiting ε-almost C-stationary points are not known in general. In the next section we will use this representation and a particular choice of discrete space to define an estimator that can be computed numerically.

4.5.2 Error estimator for finite element discretisation

The abstract error representation of the previous subsection does not depend on the discrete space. We now use the properties of the finite element space we introduced in Section 4.4.1 to define local error estimators that can be computed numerically. For simplicity we will only derive an estimator for the elliptic operator $A = -\Delta$, which satisfies the necessary assumptions.

Let $\mathcal{N}$ denote the set of interior vertices of the triangulation (i.e. vertices of T_h that are contained in Ω), and let $\mathcal{E}$ denote the set of edges of triangles in the triangulation. Let $\mathcal{N}(T)$ denote the set of vertices of a triangle $T \in T_h$ and $\mathcal{E}(T)$ denote the set of edges of T.

For a function v_h which is piecewise quadratic over the elements of T_h we can define the edge jump for $E \in \mathcal{E}$ by

$$[\nabla v_h]_E := \begin{cases} (\nabla v_h|_{T_+} - \nabla v_h|_{T_-}) \cdot \nu_{T_+,E} & E \subset \Omega \\ 0 & E \subset \partial\Omega, \end{cases}$$

where T_+ and T_- are the two triangles such that $T_+ \cap T_- = E$, and $\nu_{T,E}$ is the outer unit normal of T at E. The value of $[\nabla v_h]_E$ is independent of the permutation of T_+ and T_-. For $v_h \in V_h$ we have using integration by parts that for $z \in W_0^{1,q'}(\Omega)$,

$$\begin{aligned} a(v_h, z) &= \sum_{T \in T_h} \int_T \nabla v_h \cdot \nabla z \mathrm{d}x \\ &= \sum_{T \in T_h} \int_{\partial T} \nabla v_h \cdot \nu_{T,E} z \mathrm{d}S \\ &= \sum_{T \in T_h} \frac{1}{2} \int_{\partial T} [\nabla v_h]_{\partial T} z \mathrm{d}S, \end{aligned} \tag{4.20}$$

where $[\nabla v_h]_{\partial T}$ is defined by $([\nabla v_h]_{\partial T})|_E = [\nabla v_h]_E$.

Recall that $\xi_h, \lambda_h \in V_h^*$ are defined by (4.14) and (4.15b). However these equations do not define how they should be interpreted as functions in $W^{-1,q'}(\Omega)$. Let

$$\begin{aligned} \langle \xi_h, v \rangle_{W^{-1.q'}(\Omega)} &:= \sum_{T \in T_h} \sum_{z \in \mathcal{N}(T)} \frac{1}{N_z} \xi_{h,z} v(z) \quad \forall v \in W_0^{1,q}(\Omega), \\ \langle \lambda_h, v \rangle_{W^{-1.q'}(\Omega)} &:= \sum_{T \in T_h} \sum_{z \in \mathcal{N}(T)} \frac{1}{N_z} \lambda_{h,z} v(z) \quad \forall v \in W_0^{1,q}(\Omega), \end{aligned} \tag{4.21}$$

where $N_z := |T \in T_h : z \in \mathcal{N}(T)|$ and

$$\begin{aligned} \xi_{h,z} &:= \langle \xi_h, \varphi_z \rangle_{V_h^*}, \\ \lambda_{h,z} &:= \langle \lambda_h, \varphi_z \rangle_{V_h^*}. \end{aligned}$$

Here φ_z for $z \in \mathcal{N}$ is the nodal basis function of V_h corresponding to z (see Section 2.5.1). The advantage of these definitions is that they allow us to express duality pairings involving ξ_h and λ_h as sums of local contributions from each element of T_h, as in [Günther and Hinze, 2008]. They are consistent with the definitions as

elements of V_h^*. Note that

$$\xi_h = \sum_{z \in \mathcal{N}} \xi_{h,z} \delta_z,$$
$$\lambda_h = \sum_{z \in \mathcal{N}} \lambda_{h,z} \delta_z,$$

where $\delta_z \in W^{-1,q'}(\Omega)$ is the delta function.

In the error representation of Theorem 4.10 the right hand side depends on limiting ε-almost C-stationary points, which are not known exactly. To get around this we use the heuristic approach introduced in [Bangerth and Rannacher, 2003] to construct functions which should better approximate the stationary points. Given a function $z_h \in V_h$, on every $T \in T_h$ we approximate $z_h|_T$ by a quadratic function $\tilde{z}_{h,T} : T \mapsto \mathbb{R}$. We calculate $\tilde{z}_{h,T}$ by finding the quadratic function $\mathbb{R}^2 \mapsto \mathbb{R}$ in the basis $\{1, x_1, x_2, x_1 x_2, x_1^2, x_2^2\}$ that minimises the least square distance to z_h at a number of points (that are not necessarily contained in T), and then restricting it to a function defined over T. If the minimisation problem has more than one solution we take the one with minimum l^2 norm of the coefficient vector. We can combine the quadratic functions $\tilde{z}_{h,T}$ to get a T_h-piecewise quadratic function $\tilde{z}_h \in L^2(\Omega)$.

Note that we use a least squares approach because the problem of finding an interpolating quadratic function is more subtle than perhaps it first appears. The number of points needed to uniquely determine such a quadratic function depends on the arrangement of the points. In two dimensions 6 points such that no three lie on a straight line are sufficient. If we have 6 points and three or more lie on a straight line, it is not clear if we will have no solution or infinitely many solutions. Solving the above least squares problem amounts to a linear algebra problem which has a unique solution for any number and arrangement of points, and the solution is the unique interpolating polynomial when one exists. It therefore also allows the possibility of more easily generalising our method to higher dimensions or different types of grid.

We investigated three different ways of choosing the set of points used to fit the quadratic function for a given element T.

1. *Midpoints of edges of T and vertices of neighbouring elements that are not vertices of T.* If T is an interior element this gives 6 points to determine 6 coefficients. On triangulations with the structure we use for our test examples (see e.g. Figure 4.5), regardless of which elements are refined there is a unique quadratic function which interpolates these 6 points. Elements with edges on the boundary will give either 4 or 5 points. An interpolating quadratic

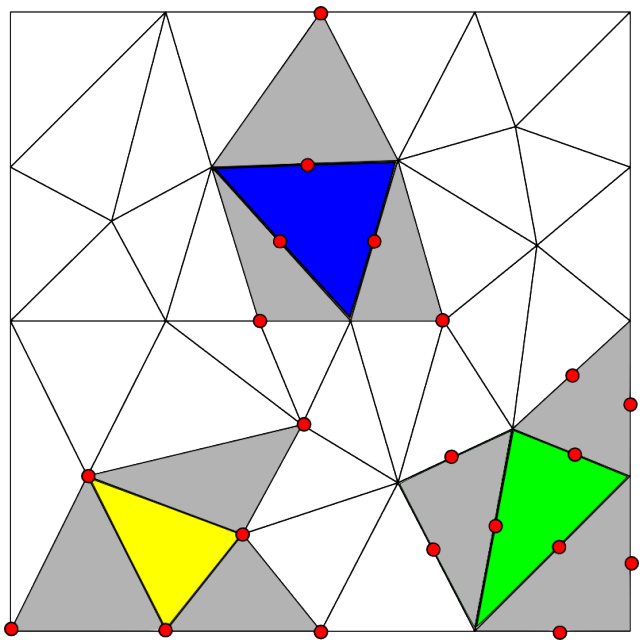


Figure 4.1: The red dots indicate the choices of points using approaches 1 (blue), 2 (yellow) and 3 (green). The grey triangles are neighbouring triangles.

function can still be found, but it is not unique and the one with coefficient vector of minimum l^2 norm is chosen.

2. *The vertices of T and neighbouring triangles.* For interior elements this gives 6 points and we can find an interpolating quadratic function. In this case the constructed functions agree exactly with the discrete functions at vertices. So some terms in the estimator that we derive are zero, even though they would likely not be zero if a limiting ε-almost C-stationary point was known exactly.

3. *The midpoints of edges of T and neighbouring triangles.* For interior elements this gives 9 points to determine 6 coefficients. As a result we generally cannot find an interpolating quadratic function.

These approaches are illustrated in Figure 4.1. We use the first approach of choosing points as exact interpolation makes it intuitively clearer what the approximate continuous solution looks like. It also does not cause potentially important terms in the estimator to be zero. This results in more refinement around the boundary of the active set, which we believe is helpful. Note that our solutions typically exhibit very sharp spikes near the point evaluations. A quadratic interpolation in this region is likely to be inaccurate.

We now use the expression in Theorem 4.10 to devise an estimator η for

$|J(y^*, u^*) - J(y_h^*, u_h^*)|$. We need the absolute value sign since feasible points for the discrete problem are not necessarily feasible for the continuous problem, so $J(y^*, u^*) - J(y_h^*, u_h^*)$ could be negative.

Let

$$\eta := \sum_{T \in T_h} \eta_T,$$

where η_T gives an indication of the local contribution to the absolute error in the objective functional caused by the discretisation at element T. To define such an η_T we decompose it into components based on the different lines of the expression in Theorem 4.10, so

$$\eta_T = \eta_{\text{PDE1},T} + \eta_{\text{PDE2},T} + \eta_{\text{CM},T}. \tag{4.22}$$

We will now define $\eta_{\text{PDE1},T}, \eta_{\text{PDE2},T}, \eta_{\text{CM},T}$.

Using (4.20), (4.21) and $\delta p_h = p_h^*$ we can rewrite (4.19a) as

$$\sum_{T \in T_h} \left(\int_{\partial T} \frac{1}{2} [\nabla y_h^*]_{\partial T} (p^* - p_h^*) \mathrm{d}S - \int_T (u_h^* + f)(p^* - p_h^*) \mathrm{d}x \right. \\ \left. - \sum_{z \in \mathcal{N}(T)} \frac{1}{N_z} \xi_h^* (p^*(z) - p_h^*(z)) \right) \tag{4.23}$$

The function p^* is unknown, so this quantity cannot be computed. Therefore we replace p^* in (4.23) by the approximate continuous solution $\tilde{p}_h^*$ to get something which is computable. This motivates us to define an estimator for the contribution to the error from (4.19a) on an element T as

$$\eta_{\text{PDE1},T} := \left| \int_{\partial T} \frac{1}{2} [\nabla y_h^*]_{\partial T} (\tilde{p}^* - p_h^*) \mathrm{d}S - \int_T (u_h^* + f)(\tilde{p}^* - p_h^*) \mathrm{d}x \right. \\ \left. - \sum_{z \in \mathcal{N}(T)} \frac{1}{N_z} \xi_{h,z}^* (\tilde{p}_h^*(z) - p_h^*(z)) \right|.$$

Note that we have chosen $\delta p_h = p_h^*$ as p_h^* should be close to p^*, and this makes it clearer that $\eta_{\text{PDE1},T}$ is small and should go to zero as $h \to 0$.

To define an estimator corresponding to line (4.19b) we follow the same approach and use (4.20), (4.21), and $\delta y_h = y_h^*$. In addition we localise the point evaluation term to each element by splitting the points into those at vertices of elements, those on edges but not vertices of elements, and those in the interior of

elements. In particular, for a continuous function ψ,

$$\sum_{\omega\in I}\psi(\omega)=\sum_{T\in T_h}\left(\sum_{\omega\in I\cap\mathrm{int}(T)}\psi(\omega)+\frac{1}{2}\sum_{\omega\in I\cap\partial T\setminus\mathcal{N}}\psi(\omega)+\sum_{\omega\in I\cap\mathcal{N}}\psi(\omega)\right).$$

This motivates us to define the estimator for the contribution to the error from (4.19b) on an element T as

$$\begin{aligned}\eta_{\mathrm{PDE2},T}:=&\frac{1}{2}\Bigg|\int_{\partial T}\frac{1}{2}[\nabla p_h^*]_{\partial T}(\tilde{y}_h^*-y_h^*)\mathrm{d}S-\sum_{z\in\mathcal{N}(T)}\frac{1}{N_z}\lambda_{h,z}^*(\tilde{y}_h^*(z)-y_h^*(z))\\&+\sum_{\omega\in I\cap\mathrm{int}(T)}(y_h^*(\omega)-g_\omega)(\tilde{y}_h^*(\omega)-y_h^*(\omega))\\&+\frac{1}{2}\sum_{\omega\in I\cap\partial T\setminus\mathcal{N}}(y_h^*(\omega)-g_\omega)(\tilde{y}_h^*(\omega)-y_h^*(\omega))\\&+\sum_{\omega\in I\cap\mathcal{N}}(y_h^*(\omega)-g_\omega)(\tilde{y}_h^*(\omega)-y_h^*(\omega))\Bigg|.\end{aligned}\tag{4.24}$$

Note that (4.19c) is zero due to (4.9c) and (4.19d) is zero from (4.15d), so we do not need estimators for these terms.

We now consider an estimator for the lines (4.19e) and (4.19f). To handle the terms with discrete multipliers paired with continuous solutions we replace y^* and p^* by the approximate continuous solutions $\tilde{y}_h^*$ and $\tilde{p}_h^*$ and note that the discrete multipliers satisfy the discrete stationarity system. So for $\langle\lambda_h^*,\tilde{y}_h^*\rangle_{W^{-1,q'}(\Omega)}$ and $\langle\xi_h^*,\tilde{p}_h^*\rangle_{W^{-1,q}(\Omega)}$ we get

$$\begin{aligned}\langle\lambda_h^*,\tilde{y}^*\rangle_{W^{-1,q'}(\Omega)}=&\langle\lambda_h^*,\tilde{y}^*-y_h^*\rangle_{W^{-1,q'}(\Omega)},\\\langle\xi_h^*,\tilde{p}^*\rangle_{W^{-1,q'}(\Omega)}=&\langle\xi_h^*,\tilde{p}^*-p_h^*\rangle_{W^{-1,q'}(\Omega)}.\end{aligned}$$

Since we used the stationarity conditions to include y_h^* and p_h^* we get an estimator that we can see should be small, as we expect the difference between the approximate continuous solution and the discrete solution to be small due to the way the former is constructed from the latter.

To handle the terms containing continuous multipliers paired with discrete solutions we have to work a little harder, as the continuous multipliers are hard to

approximate directly. We use the continuous stationarity system (4.9) to note that

$$\begin{aligned}\langle \lambda^*, y_h^* \rangle_{W^{-1,q}(\Omega)} &= \langle \lambda^*, y_h^* - y^* \rangle_{W^{-1,q'}(\Omega)} \\ &= a(y_h^* - y^*, p^*) - \sum_{\omega \in I} (y^*(\omega) - g_\omega)(y_h^*(\omega) - y^*(\omega)), \qquad (4.25) \\ \langle \xi^*, \mu_h^* \rangle_{W^{-1,q}(\Omega)} &= \langle \xi^*, p_h^* - p^* \rangle_{W^{-1,q}(\Omega)} \\ &= a(y^*, p_h^* - p^*) - (u^* + f, p_h^* - p^*)_{L^2(\Omega)}. \qquad (4.26)\end{aligned}$$

We then use the relations (4.20) and (4.21), and replace y^* and p^* by the computable approximate continuous solutions $\tilde{y}_h^*$ and $\tilde{p}_h^*$. Again we have used the stationarity conditions to introduce y^* in (4.26) and p^* in (4.25) so that it is clear the estimators should be small.

This motivates us to define the following estimator for the contribution to the error of (4.19e) and (4.19f) on an element T, consisting of a sum of estimators for the terms in (4.19):

$$\begin{aligned}\eta_{\mathrm{CM},T} :=& \frac{1}{2} \left| \sum_{z \in \mathcal{N}(T)} \frac{1}{N_z} \lambda_{h,z}^* (\tilde{y}_h^*(z) - y_h^*(z)) \right| + \frac{1}{2} \left| \sum_{z \in \mathcal{N}(T)} \frac{1}{N_z} \xi_{h,z}^* (\tilde{p}_h^*(z) - p_h^*(z)) \right| \\ &+ \frac{1}{2} \Bigg| \int_{\partial T} \frac{1}{2} [\nabla \tilde{p}_h^*]_{\partial T} (\tilde{y}_h^* - y_h^*) \mathrm{d}S - \sum_{\omega \in I \cap \mathcal{N}} (\tilde{y}_h^*(\omega) - g_\omega)(\tilde{y}_h^*(\omega) - y_h^*(\omega)) \\ &+ \sum_{\omega \in I \cap \mathrm{int}(T)} (\tilde{y}_h^*(\omega) - g_\omega)(\tilde{y}_h^*(\omega) - y_h^*(\omega)) \\ &+ \frac{1}{2} \sum_{\omega \in I \cap \partial T \setminus \mathcal{N}} (\tilde{y}_h^*(\omega) - g_\omega)(\tilde{y}_h^*(\omega) - y_h^*(\omega)) \Bigg| \\ &+ \frac{1}{2} \left| \int_{\partial T} \frac{1}{2} [\nabla \tilde{y}_h^*]_{\partial T} (\tilde{p}^* - \tilde{p}_h^*) \mathrm{d}S - \int_T (\tilde{u}_h^* + f)(\tilde{p}^* - \tilde{p}_h^*) \mathrm{d}x \right|.\end{aligned}$$

Using the above ideas it is natural to define an estimator for the active set terms (4.19g), (4.19h) as

$$\begin{aligned}\eta_{\mathrm{AS},T} :=& \frac{1}{2} \left| \int_T (u_h^* - a) \tilde{\sigma}_a^* \mathrm{d}x \right| + \frac{1}{2} \left| \int_T (\tilde{u}_h^* - a) \sigma_{a,h}^* \mathrm{d}x \right| \\ &+ \frac{1}{2} \left| \int_T (u_h^* - b) \tilde{\sigma}_b^* \mathrm{d}x \right| + \frac{1}{2} \left| \int_T (\tilde{u}_h^* - b) \sigma_{b,h}^* \mathrm{d}x \right|,\end{aligned}$$

where we have replaced $u^*, \sigma_a^*, \sigma_b^*$ by the approximations

$$\tilde{u}_h^* = \mathbb{P}_{[a,b]} \left(-\frac{1}{\nu} \tilde{u}_h^* \right), \quad \tilde{\sigma}_{a,h}^* = \max(0, a + \frac{1}{\nu} \tilde{p}^*), \quad \tilde{\sigma}_{b,h} = \max(0, -\frac{1}{\nu} \tilde{p}^* - b).$$

We have now defined an estimator corresponding to every nonzero component of (4.19), and so have completed the definition of the fully a posteriori local estimator (4.22).

Remark 4.11. *Note the absolute value signs around the terms in $\eta_{PDE1,T}, \eta_{PDE2,T}$ and $\eta_{CM,T}$. There are other possibilities here. For example, we could have left out the absolute values and instead defined $\eta_T = |\eta_{PDE1,T} + \eta_{PDE2,T} + \eta_{CM,T}|$, but then errors in one quantity (e.g. from replacing the exact stationary point with an approximation) could cancel with errors in another quantity.*

4.6 Finite element scheme

In this section we describe our scheme for solving the discrete stationarity system (4.15). Motivated by Theorem 4.7 we find an approximate solution to the discrete stationarity system by solving the penalised stationarity system (4.10) for large γ. We can solve the penalised stationarity system using a Newton-type method, but we find that the radius of convergence decreases as γ increases. To get around this we solve the penalised stationarity system for a small γ, then use this solution as an initial guess in the Newton-type method to solve it for a slightly larger γ. We repeat this process until the penalisation γ is sufficiently large that solutions of the penalised stationarity system almost satisfy the discrete stationarity system.

4.6.1 Solving the discrete penalised stationarity system

In a similar way to the previous chapters, to solve the discrete penalised stationarity system (4.10) we find $(y_h, p_h) \in V_h \times V_h$ such that

$$\begin{pmatrix} a(y_h, v_h) - (-\frac{1}{\nu}p_h + (a + \frac{1}{\nu}p_h)^+ - (-\frac{1}{\nu}p_h - b)^+ - f + \gamma \max_{\varepsilon(\gamma)}(0, -y_h), v_h) \\ a(v_h, p_h) + \sum_{\omega \in I}(y_h(\omega) - g_\omega)w_h(\omega) + (\gamma \max'_{\varepsilon(\gamma)}(0, -y_h)p_h, w_h) \end{pmatrix} = 0$$

for all $v_h, w_h \in V_h$, where v^+ denotes the nonnegative part $max(0, v)$ of a function v. We can then determine u_h using the relation $u_h = \mathbb{P}_{[a,b]}\big(-\frac{1}{\nu}p_h\big)$. In order to solve this we let $F_h^\gamma : V_h \times V_h \to V_h^* \times V_h^*$ with $F_h^\gamma(y_h, p_h)(v_h, w_h)$ defined by the left hand side of the above system, then apply the semismooth Newton method to find (y_h, p_h) such that

$$F_h^\gamma(y_h, p_h) = 0 \text{ in } V_h^* \times V_h^*.$$

See Section 2.5.1 for more details on the semismooth Newton method and its implementation.

Note that in contrast to Chapter 2 and Chapter 3, this problem does not necessarily have a unique solution. This is not a problem for Newton-type methods, as the problem we solve at each iteration does have a unique solution. We also no longer expect local superlinear convergence of the semismooth Newton method, but it works sufficiently well in practice so long as we globalise it. In particular we use a simple Armijo-type backtracking, which is shown in Algorithm 2. This introduces parameters $\tau_0 > 0$ and $\zeta, \kappa \in (0, 1)$. We found that the parameter values $\tau_0 = 1e-8, \zeta = \kappa = 0.5$ work well. For the coupling between γ and ε we take $\varepsilon(\gamma) = 0.1 * \gamma^{-1.2}$, which satisfies Assumption 4.2.

Algorithm 2 For solving the discrete penalised stationarity system (2)

function SOLVEPEN($T_h, \gamma, y_h^0, p_h^0$, DATA) ▷ DATA $= (\Omega, \nu, f, a, b, I, \{g_\omega\}_{\omega \in I})$
  $k \leftarrow 0$
  while $\|F_h^\gamma(y_h^k, p_h^k)\|_{H^{-1}} > \delta$ **do** ▷ $\delta = 1e-8$
    Compute $(\delta y, \delta p)$ by solving $(F_h^\gamma)'(y_h^k, p_h^k)(\delta y, \delta p) = -F_h^\gamma(y_h^k, p_h^k)$.
    $(y_h^{k+1}, p_h^{k+1}) \leftarrow (\delta y, \delta p)$
    $\tau \leftarrow 1$
    while $\|F_h^\gamma(y_h^{k+1}, p_h^{k+1})\|_{H^{-1}} > (1 - \kappa\tau)\|F_h^\gamma(y_h^k, p_h^k)\|_{H^{-1}}$ **and** $\tau > \tau_0$ **do**
      $\tau \leftarrow \zeta \cdot \tau$
      $(y_h^{k+1}, p_h^{k+1}) \leftarrow (y_h^k, p_h^k) + \tau(\delta y, \delta p)$
    end while
    $k \leftarrow k + 1$
  end while
  return y_h^k, p_h^k
end function

Mesh independence

Although the analytical results for the function space semismooth Newton method do not hold when solving the discrete penalised stationarity system (4.10), we nevertheless observe good mesh independence properties. Solving this with $\gamma = 100$ for Example 1 (introduced in Section 4.7.1) we get Table 4.1. The number of Newton iterations needed varies, but does not increase much as h is decreased.

We do not get quadratic convergence of the Newton method, but the convergence is on average superlinear (but again quite variable). See Table 4.2 for experimentally observed convergence rates as defined in (2.62) for Example 4.12 with $\gamma = 100$ and $h = 0.0110485$.

h	# iterations
0.0883883	15
0.0441942	14
0.0220971	13
0.0110485	13
0.00552427	18

Table 4.1: Number of Newton iterations needed for a given h.

iteration k	$\|F_h^k(u_h^\gamma)\|_{H^{-1}(\Omega)}$	EOC_k
0	0.00637888	-
1	0.00134885	0.49466819
2	0.000625419	1.5117777
3	0.000195682	0.14852672
4	0.000164665	0.91598973
5	0.000140588	16.172142
6	1.09064e-05	0.012929208
7	1.05518e-05	32.041547
8	3.6591e-06	1.3864263
9	8.42722e-07	1.2946135
10	1.25928e-07	0.25735196
11	7.7208e-08	3.7506911
12	1.23251e-08	0.31283993
13	6.9422e-09	-

Table 4.2: Convergence rate of Newton method.

4.6.2 Solving the discrete stationarity system

We use Algorithm 3 for solving the discrete stationarity system (4.15). The algorithm makes use of the function SOLVEPEN, which is defined by Algorithm 2, to solve the discrete penalised stationarity system.

The function RESIDUAL computes a residual r to determine how well the solution to the discrete penalised stationarity system solves the discrete stationarity system. When r is smaller than a tolerance δ we consider the system sufficiently well solved and terminate the algorithm. The primal and dual equations are solved almost exactly by the linear solver, so the component of r corresponding to these parts of the optimality system is negligible. Choosing components of r corresponding to the rest of the optimality system is harder, as there are many possible choices. We could compute quantities inspired by the finite dimensional MPEC, as in [Hintermüller et al., 2013]. For our implementation it is more natural to compute function space based residuals. An advantage of this is that we can calculate quantities which take into account the size of the region over which y_h is negative, such as $\| \max(0, -y_h)\|_{L^2}$. We take $r = r_1 + r_2 + r_3$ with

$$
\begin{aligned}
r_1 &= |\langle \xi_h, p_h \rangle| \\
r_2 &= |\langle \xi_h, y_h \rangle| + \sum_{i \in \mathcal{N}} \max(0, -\langle \xi_h, \phi_i \rangle) + \| \max(0, -y_h)\|_{L^2} \\
r_3 &= |\langle \lambda_h, y_h \rangle|.
\end{aligned}
$$

There may be better choices for the component of the residual testing whether $\xi \geq 0$. However in all the examples we tested, the contribution from any sensibly defined residual for this term was dominated by the other components of the residual.

The penalisation parameter γ is increased by multiplying it by the update factor $\delta\gamma$. We take an initial γ of $\gamma_{\min} = 100$, for which the semismooth Newton method converged for all the examples we tested. Taking a large value for $\delta\gamma$ does not significantly decrease the total number of Newton iterations needed, as then the initial guess is only just inside the radius of convergence of the Newton method, and convergence is slow initially. Moreover, if $\delta\gamma$ is too large then we may have no convergence at all. Therefore we take a moderate value of $\delta\gamma = 1.2$, which also worked for all examples we tested.

Algorithm 3 For solving the discrete stationarity system (4.15)

function SOLVE(T_h, y_h^0, p_h^0, DATA) ▷ DATA $= (\Omega, \nu, f, a, b, I, \{g_\omega\}_{\omega \in I})$
  $\gamma \leftarrow \gamma_{\min}, \gamma_- \leftarrow 0$
  loop
    $(y_h^\gamma, u_h^\gamma) \leftarrow$ SOLVEPEN($T_h, \gamma, y_h^{\gamma_-}, p_h^{\gamma_-}$, DATA)
    Compute $\xi_h^\gamma, \lambda_h^\gamma$ according to (4.14), (4.15b)
    if RESIDUAL($y_h^\gamma, p_h^\gamma, \xi_h^\gamma, \lambda_h^\gamma$) $> \delta$ **then** ▷ $\delta = 1e-5$
      return $y_h^\gamma, p_h^\gamma, \xi_h^\gamma, \lambda_h^\gamma$
    end if
    $\gamma_- \leftarrow \gamma, \gamma \leftarrow \delta\gamma \cdot \gamma$
  end loop
end function

4.7 Numerical results

We test two different approaches for solving the optimal control problem numerically to a high degree of accuracy. The first will solve the discrete penalised problem on a sequence of uniformly refined grids for increasing γ. The second will use the a posteriori error estimator we derived in Section 4.5 to adaptively refine the triangulation.

4.7.1 Uniform refinement

In order to compute functions which closely approximate limiting ε-almost C-stationary points, we could use Algorithm 3 to solve the discrete penalised stationarity system (4.15) on a fine triangulation for increasingly large γ. This is computationally expensive, as a solves on fine triangulations is required for each choice of γ.

An alternative approach, which we state in Algorithm 4, is to use Algorithm 2 and increase γ at the same time as uniformly refining the triangulation. The motivation for this is the observation that h is sufficiently small by the time γ is large that (y_h^γ, p_h^γ) is always inside of the radius of convergence of the semismooth Newton method. The downside of the approach is that stationarity conditions such as $y_h \geq 0$ may not be satisfied for a given h.

Note that solving the discrete penalised stationarity system for a large γ then uniformly refining the triangulation is not a viable strategy. In this case (y_h, p_h) is not a sufficiently good initial value for the Newton method without decreasing γ to a smaller value.

Algorithm 4 uses the function SOLVEPEN from Algorithm 2 to solve the discrete penalised stationarity system (4.15). The function REFINEALL outputs a uniform refinement of the current triangulation as well as the usual prolongations of

Algorithm 4 Uniform refinement

Input: T_h, y_h^0, p_h^0 and $h_{\min}, \gamma_{\min} > 0$

DATA $\leftarrow (\Omega, \nu, f, a, b, I, \{g_\omega\}_{\omega \in I})$
$\gamma_- \leftarrow 0$
loop
  $\gamma \leftarrow \gamma_{\min}$
  while $\gamma > ch^\alpha$ **do** ▷ $c = 10, \alpha = 2$
    $(y_h^\gamma, p_h^\gamma) \leftarrow$ SOLVEPEN$(T_h, \gamma, y_h^{\gamma_-}, p_h^{\gamma_-}, \text{DATA})$
    $\gamma_- \leftarrow \gamma, \gamma \leftarrow \delta\gamma \cdot \gamma$ ▷ $\delta\gamma = 1.2$
  end while
  if $h > h_{\min}$ **then break**
  end if
  $(T_h, y_h^\gamma, p_h^\gamma) \leftarrow$ REFINEALL$(T_h, y_h^\gamma, p_h^\gamma)$
end loop
Compute $\xi_h^\gamma, \lambda_h^\gamma$ according to (4.14), (4.15b)

Output: $y_h^\gamma, p_h^\gamma, \xi_h^\gamma, \lambda_h^\gamma$

the discrete functions so that that they are defined over the refined triangulation. In particular, it bisects all the triangular elements in such a way that the triangulation remains conforming.

We use the same value of $\delta\gamma$ as in Algorithm 2 and use numerical experiments to calibrate α.

Calibrating α

Denote by u an ε-almost limiting C-stationary point, by u_h a solution of the discrete stationarity system (4.15), and by u_h^γ a solution of the discrete penalised stationarity system (4.10).

We aim to find an exponent α such that for $\gamma(h) = Ch^\alpha$ we get the highest order convergence of $\|u - u_h^{\gamma(h)}\|_{L^2(\Omega)}$ with respect to h. We do this by noting that

$$\|u - u_h^{\gamma(h)}\|_{L^2(\Omega)} \leq \|u - u_h\|_{L^2(\Omega)} + \|u_h - u_h^{\gamma(h)}\|_{L^2(\Omega)},$$

where the first term can be thought of as discretisation error and the second term can be thought of as penalisation error. So if we can find α_1 and α_2 such that

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{\alpha_1}, \quad \|u_h - u_h^\gamma\|_{L^2(\Omega)} \leq C\gamma^{-\alpha_2},$$

with C independent of h in both inequalities, then taking $\gamma(h) = Ch^{-\alpha_1/\alpha_2}$ will give that

$$\|u - u_h^{\gamma(h)}\|_{L^2(\Omega)} = O(h^{\alpha_1})$$

i.e. we should take $\alpha = -\alpha_1/\alpha_2$.

We will calibrate α_1 and α_2 using the following example. A discrete stationarity point for this example is visualised in Figure 4.4.

Example 4.12. *Let* $\Omega = (0,1)^2$*,* $A = -\Delta$*,* $\nu = 0.003$*,* $f = 0$*,* $a = -100$*,* $b = 100$*,*

$$I = \{(0.125, 0.125), (0.125, 0.5), (0.375, 0.375).(0.5, 0.125)\},$$

and $g_\omega = 1$ *at* $(0.125, 0.125)$ *and* $g_\omega = 0$ *otherwise.*

First note that $\|u_h - u_{h'}\|_{L^2(\Omega)}$ is a reasonable approximation to $\|u_h - u\|_{L^2(\Omega)}$ for $h' \ll h$. Calculating this quantity with $h' = 0.00552427$ gives Figure 4.2(a). Based on this we speculate that

$$\|u_h - u\|_{L^2(\Omega)} \leq Ch^{\alpha_1},$$

with $\alpha_1 \approx 1$ and C independent of h.

Next we calculate compute $\|u_h^\gamma - u_h\|_{L^2(\Omega)}$, making sure that $\gamma \ll \gamma' = 147789$, the final γ in Algorithm 3. For $h = 0.0110485$ this gives Figure 4.2(b), and we speculate that

$$\|u_{\gamma,h} - u_h\|_{L^2(\Omega)} \leq C\gamma^{-\alpha_2},$$

with $\alpha_2 \in (1, 2)$. Increasing γ is cheaper than decreasing h so we are conservative and suppose $\alpha_2 \approx 1$. As we have the function space convergence result Theorem 4.7, C should be roughly independent of γ. Combining these two experimentally observed convergence relationships we get that in order to stop the penalisation error dominating the discretisation error we should take $\gamma(h) = \frac{c}{h}$ for some constant c.

We finish this section by testing our calibrated method on a simple example for which we know the exact solution, allowing the quantities $\|u_{\gamma,h} - \tilde{u}\|_{L^2(\Omega)}$ and $\left|\hat{J}(\tilde{u}) - \hat{J}_h(u_h)\right|$ to be computed exactly. Taking $c = 10$ in this relationship gives Figure 4.3.

We finish this section by testing the uniform refinement algorithm on a simple problem for which the solution is known exactly.

Example 4.13. *Let* $\Omega = (-1, 1)^2$ *and take*

$$y(x) = \begin{cases} \frac{1}{2}(\cos(\frac{3}{2}\pi\,|x|) + 1) & |x| < \frac{2}{3}, \\ 0 & |x| \geq \frac{2}{3}. \end{cases}$$

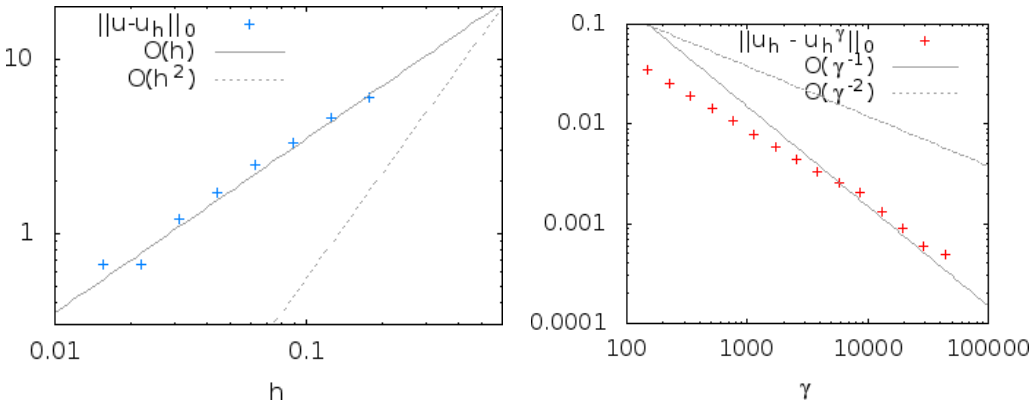


Figure 4.2: Convergence with respect to h and γ for Example 4.12.

Then y is the solution to variational inequality (4.1) with $A = -\Delta$ for $f = -\Delta y$ and $u = 0$. Now take $\nu = 1$, $b = a = -\infty$ and $g_\omega = y(\omega)$ for all $\omega \in I$. For any selection of points I the global minimiser of J is $(y(0), 0)$, as this is feasible and $J(y, 0) = 0$. We will take

$$I := \{(0,0), (\pm 0.5, \pm 0.5), (0, \pm 0.5), (\pm 0.5, 0)\}.$$

We observe in Figure 4.3 the expected $\mathcal{O}(h)$ convergence of $\|u_{\gamma,h} - \tilde{u}\|_{L^2(\Omega)}$ and also $\mathcal{O}(h^2)$ convergence of $\left|\hat{J}(\tilde{u}) - \hat{J}_h(u_h)\right|$, where $\tilde{u}$ is a discrete stationary point for small h.

4.7.2 Adaptive refinement

An alternative to uniform refinement is to use an adaptive finite element method (AFEM) guided by the estimator we derived in Section 4.5.2. This estimator guides the refinement so that less computational effort is needed to find a solution to the discrete stationarity system such that $|J(u_h, y_h) - J(u, y)|$ is small.

The AFEM can be summarised as follows, and is outlined more precisely in Algorithm 5: Starting with a coarse triangulation, solve (4.15) on the current triangulation (as described in Section 4.6.2) then compute the local error indicator (4.22) for each element T. Mark elements with large local error indicators for refinement, and refine a superset of these. Additional elements are refined in order to keep the triangulation conforming. The process can then be repeated by solving (4.15) again on the new triangulation. The steps of *solve*, *estimate*, *mark* and *refine* continue until some stopping criterion is met, such as the estimator is sufficiently small, or

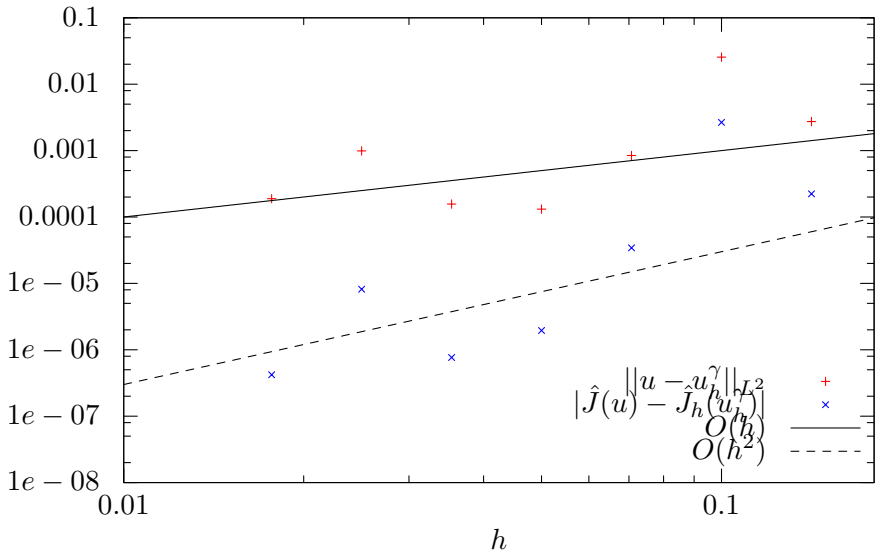


Figure 4.3: Convergence of u_h^γ for Example 4.13.

the complexity of the solve step reaches a certain level.

Algorithm 5 AFEM for MPEC

Input: T_h, y_h^0, p_h^0 and $N > 0$

DATA $\leftarrow (\Omega, \nu, f, a, b, I, \{g_\omega\}_{\omega \in I})$
$(y_h, p_h) \leftarrow (y_h^0, p_h^0)$
loop
  $(y_h, p_h, \xi_h, \lambda_h) \leftarrow$ SOLVE$(T_h, y_h, p_h,$ DATA$)$
  $\{\eta_T\}_{T \in T_h} \leftarrow$ ESTIMATE$(y_h, p_h, \xi_h, \lambda_h)$
  if $|\mathcal{N}| > N$ **then break**
  end if
  $\mathcal{M}_h \leftarrow$ MARK$(T_h, \{\eta_T\}_{T \in T_h}, \theta)$ ▷ $\theta = 0.3$
  $(T_h, y_h, p_h) \leftarrow$ REFINE$(T_h, \mathcal{M}_h, y_h, p_h)$
end loop

Output: $y_h, p_h, \xi_h, \lambda_h$

As we observed in the previous section, solutions of the discrete stationarity system on a given triangulation are in general not good initial values for the Newton method for solving on refined triangulations. So in contrast to the uniform refinement approach in Algorithm 4, after refining the triangulation we need to drop γ back to a small value. This is not a serious drawback as we have already saved computation time by carefully placing degrees of freedom only where they are needed.

The function SOLVE uses Algorithm 2 to solve (4.15). The function ESTIMATE

calculates the estimator defined in Section 4.5.2 from the discrete solution. The function MARK determines the set of elements $\mathcal{M}_h$ to be refined. In particular, we take $\mathcal{M}_h$ to be the set of minimal cardinality such that for a parameter $\theta \in (0, 1)$,

$$\theta\eta_h \leq \sum_{T\in\mathcal{M}_h} \eta_T.$$

Larger values of θ lead to inefficiency in the refinement, and smaller values of θ allow the estimator to be skewed by inaccurate values of the local estimators near the point evaluations, which are caused by bad properties of the approximate discrete solution at spikes. The drawback of this marking strategy is that it is hard to implement in parallel. When doing large runs in parallel we mark elements for refinement if

$$\eta_T > \frac{0.9\tau^2}{|T|} \sum_{T\in T_h} \eta_T,$$

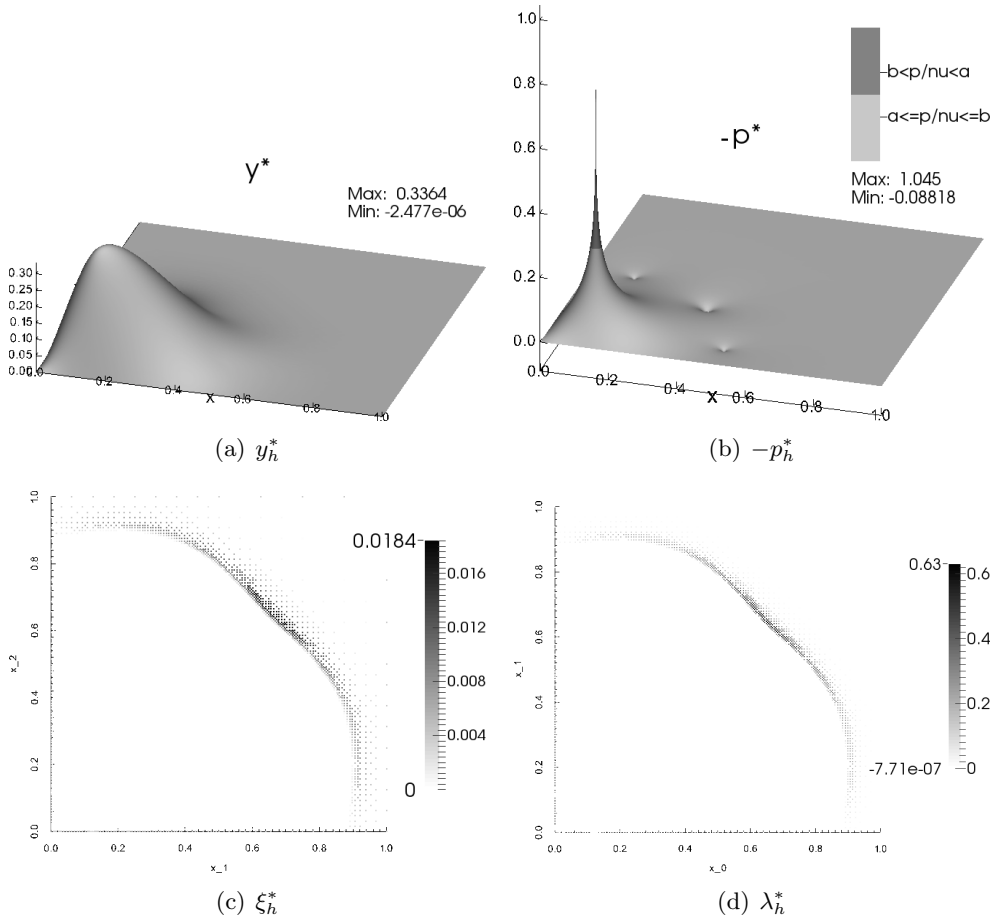
where τ is a small number that heuristically is the value of the estimator that we would like to achieve. The sequence of refinements produced by this method are more easily skewed by inaccurate values of local estimators, so it does not perform so well in practice.

The function REFINE refines all marked elements as well as additional elements in order to ensure the triangulation remains conforming. The refinement rule we use bisects the triangle shaped elements by inserting a new vertex on the longest edge. In addition this function prolongs the discrete solutions in the expected way so they are defined over the refined triangulation, ready to be used as an initial value for the next AFEM loop.

We now show the numerical results from applying the AFEM to two examples. Figure 4.4 shows a solution of the discrete stationary system for Example 4.12 computed on a triangulation with 96448 degrees of freedom (DoFs) that has undergone 15 levels of adaptive refinement and two additional uniform refinements. Observe that the state is influenced to track the prescribed values: $y_h^*(0.125.0.125) \approx 0.3$, $y_h^*(0.125, 0.5) = y_h^*(0.5, 0.125) \approx 0.04$, $y_h^*(0.375, 0.375) \approx 0.07$. However the control constraint $u_h \leq b$ is active (shown by the dark part of the plot of p_h^*), preventing y_h^* getting closer to the prescribed value of 1 at $\omega = (0.125, 0.125)$. We see that this example has a biactive set $\{y_h = 0\} \cap \{\xi_h = 0\}$ with positive measure so strict complementarity does not hold. Such problems are typically hard to solve because the active constraint gradients at the solution are linearly dependent (see e.g. the comments in [Hintermüller and Kopacka, 2009]).

Figure 4.5 shows part of the sequence of adaptively refined triangulations.

Observe that the refinement mostly takes place at the boundary between the inactive set $I := \{a < u_h < b\}$, active set $\Omega \setminus I$, and biactive set.



(a) y_h^* (b) $-p_h^*$

(c) ξ_h^* (d) λ_h^*

Figure 4.4: A solution to the discrete stationarity conditions for Example 4.12 computed using the AFEM algorithm.

We do not know the exact solution to Example 4.12, so for evaluating the effectiveness of our method we approximate it by solving the discrete stationarity conditions on a grid resulting from two additional uniform refinements of the finest adaptive grid. We denote the value of the objective functional evaluated at the solutions computed on this grid by J^*. We denote the value of the objective functional evaluated on an adaptive grid by J^A and on a uniform grid by J^U. Figure 4.6 shows the error $\left|J^A - J^*\right|$ and the estimator η^A on adaptively refined grids, and the error $\left|J^U - J^*\right|$ on uniformly refined grids. We see that the AFEM gives lower errors for a given number of degrees of freedom, which is linked to computational cost. In this example we also see that the estimator is reliable (i.e. $\left|J^A - J^*\right| \leq \eta^A$) and efficient

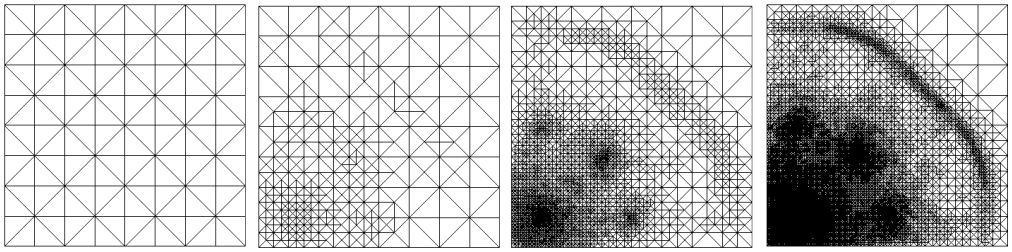


Figure 4.5: The refined triangulations for Example 4.12 at levels 0, 5, 10 and 15.

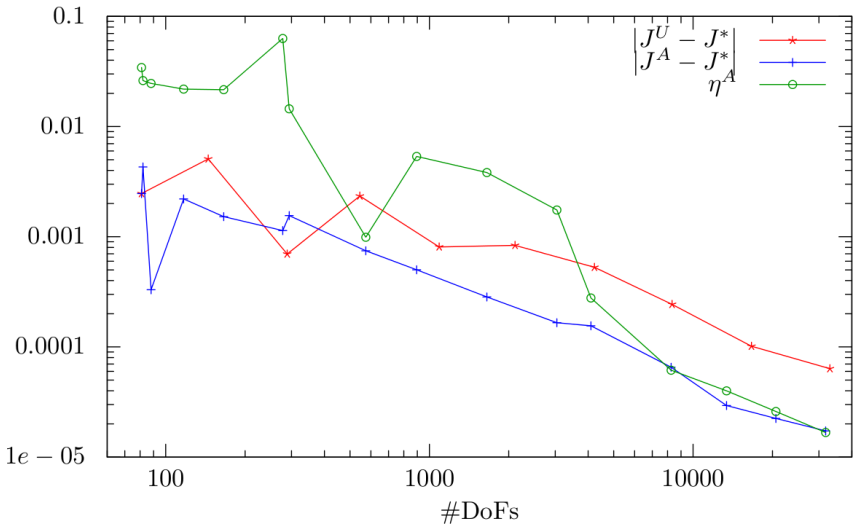


Figure 4.6: Convergence of the AFEM for Example 4.12.

(i.e. there exist $c_0, c_1 > 0$ such that $c_0 \leq \frac{\eta^A}{|J^A - J^*|} \leq c_1$).

Example 4.14. *Let* $\Omega = (-1,1)^2 \setminus [0,1]^2$ *be an 'L' shaped domain,* $b = -a = \infty$ *(i.e. no control constraints),* $A = -\Delta$*,*

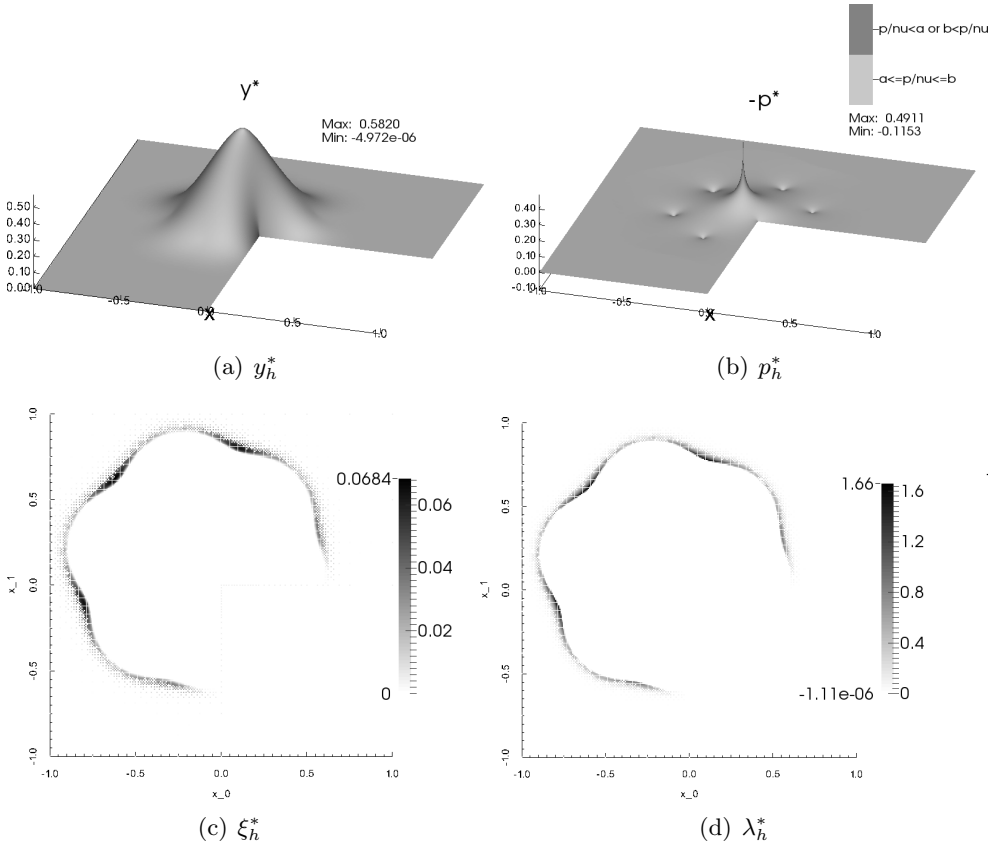
$$I = \{(-0.125, 0.125), (-0.25, -0.25), (0, 0.5), (-0.5, 0), (0.25, 0.25), (-0.375, 0.375)\},$$

and $g_\omega = 1$ *for* $\omega = (-0.125, 0.125)$ *and* $g_\omega = 0$ *otherwise. Take* $\nu = 0.003$ *and* $f = 0$ *as in Example 4.12.*

Figure 4.7 show a solution to the discrete stationarity conditions for Example 4.14 computed a triangulation with 57398 DoFs that had 9 adaptive refinements and two additional uniform refinements. We see that this problem also has a biactive set with positive measure. As there are no control constraints u_h^* is unbounded, and the state is able to get closer to the desired values: $y_h^*(-0.125, 0.125) \approx 0.6$, $y_h^*(-0.25, -0.25) = y_h^*(0.25, 0.25) \approx 0.06$, $y_h^*(0, 0.5) = y_h^*(-0.5, 0) \approx 0.08$,

$y_h^*(-0.375, 0.375) \approx 0.1$. A selection of the adaptively refined grids can be seen in Figure 4.8.

Figure 4.9 shows that the AFEM offers not only a lower error but also faster convergence because of the steepness of the solution around $(0, 0)$. So the estimator is reliable for this example but not efficient.



(a) y_h^* (b) p_h^*

(c) ξ_h^* (d) λ_h^*

Figure 4.7: A solution to the discrete stationarity conditions for Example 4.14 computed using the AFEM algorithm.

4.7.3 Optimal control of variational inequalities on curves

We finish this chapter by showing the numerical solution to Example 3.21 but with the elliptic PDE state equation replaced by an elliptic variational inequality (with $A = -\Delta$). We will not state this problem precisely or discuss the analysis or numerical analysis, as it follows in a straightforward manner by combining the ideas in Chapter 3 with the ideas in this chapter. The solution can be seen in Figure 4.10.

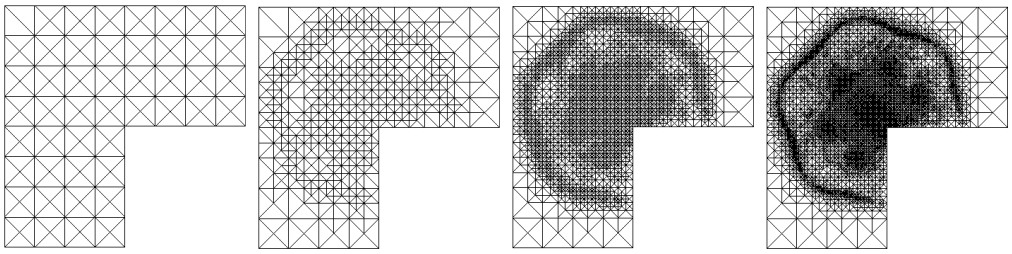


Figure 4.8: The refined triangulations for Example 4.14 at levels 0, 3, 6 and 9.

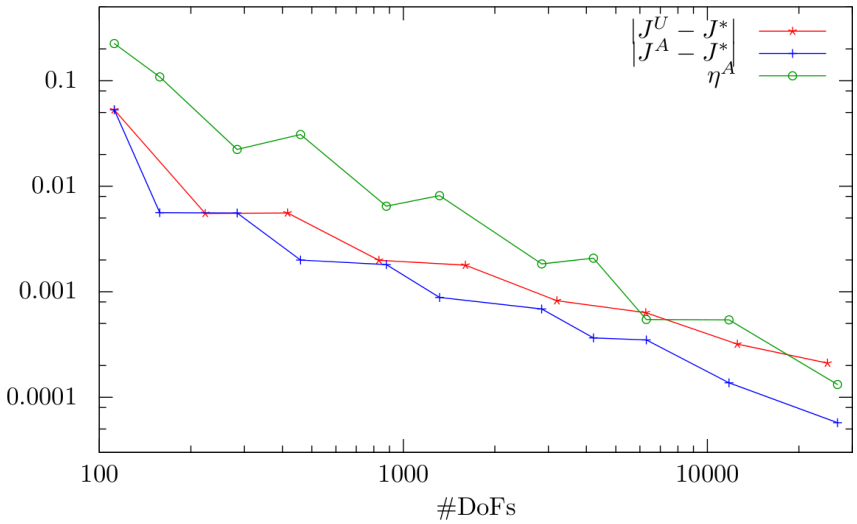
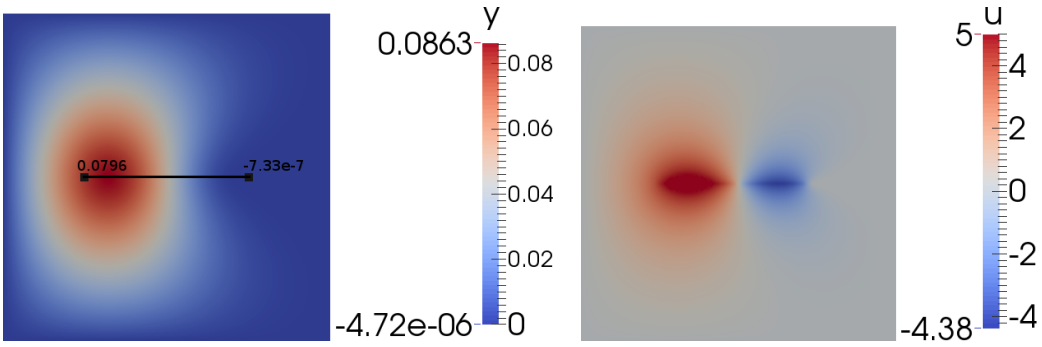


Figure 4.9: Convergence of the AFEM for Example 4.14.



(a) y^* (b) u^*

Figure 4.10: An example of optimal control of a variational inequality on a line.

Chapter 5

Phase field methods for binary recovery

5.1 Introduction

A fundamental problem in the field of image processing is the following. Suppose we have a function $\bar{u}$ defined on a bounded and piecewise smooth domain $\Omega \subset \mathbb{R}^n$ for $n \leq 3$, which has been transformed by a linear operator S, and then corrupted by additive noise ζ, such that we have data

$$g_d := S\bar{u} + \zeta.$$

The problem is to recover $\bar{u}$ given g_d. Two immediate issues are that (a) ζ is unknown, so we will not be able to find $\bar{u}$ even with a good model for the space in which it lies (b) inverting S may be ill-posed, making it difficult to find an approximation to $\bar{u}$ even if $\zeta = 0$.

We investigate this problem in the case that $\bar{u}$ is a binary function. We develop the theory with S an abstract operator, but in examples we take S to be the solution operator of an elliptic PDE. In this case the problem becomes one in PDE constrained optimal control.

Our approach to modelling the problem is to minimise an energy functional consisting of an L^2 fidelity term plus a phase field approximation to minimal perimeter regularisation. This can be thought of as a relaxation of the Mumford-Shah segmentation model. In our phase field approximation we use the Ginzburg-Landau functional with both the smooth double well and double obstacle potentials.

5.1.1 Motivating examples

First we give examples from both image processing and optimal control of PDEs motivating the study of this problem:

- **Image segmentation** - We can represent a barcode by a 1D function which takes the value -1 when the barcode is white and 1 when it is black. When a barcode is scanned by a barcode reader this function becomes blurred (due to scattering in the air) and noisy (due to measurement error and imperfections in the barcode). So the machine only sees a corrupted signal, but from this it needs to determine the scanned barcode.

- **Elliptic source recovery** - Suppose we have noisy data of a quantity y, which is related to another quantity $\bar{u}$ by some physical law. For example, let $\bar{u}$ represent a heat source, then the long term temperature distribution y may be related to $\bar{u}$ by the solution of an elliptic PDE. Our goal could be to find the heat source that produces a particular temperature distribution.

5.1.2 Background material

For the above problems to be tractable we naturally require some knowledge of the form of the operator S and the noise ζ. We also usually assume a specific form of $\bar{u}$, as this influences the best model to use. For example, in the barcode problem we could assume that the function we are trying to recover is a binary function taking the values -1 and 1, and that the bars have a minimum width. Some sets of assumptions on S, ζ and $\bar{u}$ that are made in the literature are the following:

1. *Denoising and deblurring* - S is a blurring operator (maybe the identity), ζ is Gaussian noise, and $\bar{u}$ is a piecewise smooth function ([Rudin et al., 1992], [Chambolle and Lions, 1997], [Chan and Esedoglu, 2005]).

2. *Segmentation* - S is a blurring operator (maybe the identity), ζ is Gaussian noise, and $\bar{u}$ is binary function ([Mumford and Shah, 1989], [Esedoglu, 2004], [Choksi and Gennip, 2010]). These are the assumptions we make in this work.

3. *Binary image restoration* - S is the identity, we have 'salt and pepper' noise, and $\bar{u}$ is a binary function ([Chan et al., 2006]). This kind of noise gives each point of a binary function a probability of switching to the other value, so the data g_d is also binary.

Note that the above sets of assumptions have been named using terminology from image processing. Although our problem can be thought of as either an image

processing or PDE constrained optimal control problem depending on the choice of S, we found most of the relevant literature to be from the image processing community. This is unsurprising since image processing is one of the main applications of binary recovery. We end up taking S to be the solution operator of an elliptic PDE, but try to use neutral language which reflects that our problem arises in these two fields.

For segmentation, which we focus on in this work, a large proportion of the literature modifies one of the following two models when formulating the problem of Section 5.1 mathematically. We now introduce these models so the reader can see how our approach fits with the existing literature.

- **Model 1 (Mumford-Shah).** This model, which was introduced in [Mumford and Shah, 1989], looks for piecewise smooth functions that minimise an energy functional.

  Let Ω_i be disjoint open subsets with piecewise smooth boundaries such that the closure of $\bigcup \Omega_i$ is Ω. Let u be a function that is differentiable on $\bigcup \Omega_i$, but which is allowed to be discontinuous across $\Gamma := \bigcup \partial\Omega_i \setminus \partial\Omega$. Then the Mumford-Shah model involves minimising

  $$E_1(u, \Gamma) = \frac{1}{2} \int_\Omega (u - g_d)^2 + \mu \int_{\Omega \setminus \Gamma} |\nabla u|^2 + \sigma \, |\Gamma| \,, \tag{5.1}$$

  where $|\Gamma|$ denotes the $n-1$ dimensional Hausdorff measure of Γ. The $|\Gamma|$ term encourages minimising the length of the interface over which u is discontinuous.

  If we restrict to minimising over binary functions that take the unknown value a_i on Ω_i $(i = 0, 1)$, then this energy functional becomes

  $$E_2(\{a_i\}, \Gamma) = \frac{1}{2} \sum_i \int_{\Omega_i} (a_i - g_d)^2 + \sigma \, |\Gamma| \,.$$

  For fixed Γ note that E_2 is minimised with respect to $\{a_i\}$ by setting

  $$a_i = \frac{1}{|\Omega_i|} \int_{\Omega_i} g_d.$$

  So the problem reduces to just finding Γ, the locations of the discontinuities.

  Due to the spaces of functions we are minimising over, both of the above variants of the Mumford-Shah model are nonconvex problems. In our work will use a relaxation of (5.1) based on a phase field approximation.

- **Model 2 (ROF).** The ROF (Rudin-Osher-Fatemi) model of [Rudin et al., 1992] involves solving the following constrained minimisation problem over a suitable space of functions:

$$\begin{aligned} &\text{Minimise } |u|_{TV} \\ &\text{with } \int_\Omega u = \int_\Omega g_d \text{ and } \int_\Omega (u - g_d)^2 = s^2. \end{aligned} \tag{5.2}$$

The term $|u|_{TV}$ represents the total variation of u, and it can be defined even if u is not continuous; the total variation of a function $u \in L^1(\Omega)$ is

$$|u|_{TV} := \sup\{-\int_\Omega u \mathrm{div}(\phi)\,\mathrm{d}x : \Phi \in C_c^\infty(\Omega, \mathbb{R}^n), \|\phi\|_{L^\infty(\Omega)} \leq 1\}.$$

So the total variation of a function is the total variation of the measure Du (see Section 2.1), where $\mu = Du$ is a signed measure with finite total variation defined by

$$-\int_\Omega u \mathrm{div}(\phi)\,\mathrm{d}x = \int_\Omega \phi\,\mathrm{d}\mu \quad \forall \phi\, C_c^\infty(\Omega)$$

(see e.g. [Ambrosio et al., 2000]). Sometimes the notation $\int_\Omega |\nabla u|$ is used instead of $|u|_{TV}$ to highlight that the total variation of u is equal to this quantity when it is well defined. Under the assumption that $g_d = u + \zeta$ (where ζ is the noise), the first constraint in (5.2) says that the noise has mean zero and the second constraint says that it has variance s^2.

$BV(\Omega, \mathbb{R})$ is the subspace of functions in $L^1(\Omega)$ which have finite total variation. Minimising this model over $u \in BV(\Omega, \mathbb{R})$ can be related to the following problem for some value of σ:

$$\text{Minimise } \quad \frac{1}{2}\|u - g_d\|^2_{L^2(\Omega)} + \sigma\,|u|_{TV} \text{ over } BV(\Omega, \mathbb{R}). \tag{5.3}$$

Note that (5.3) can be thought of as a relaxation of (5.1) with $\mu = 0$; we minimise over a larger space of functions in order to get a convex problem.

If we restrict to minimising over binary functions then (5.3) becomes similar to the Mumford-Shah model. Suppose u only takes the known values a_0 and a_1 with $a_0 < a_1$ (i.e. $u \in BV(\Omega, \{a_0, a_1\})$), then

$$|u|_{TV} = (a_1 - a_0)\mathrm{Per}(\{u = a_1\}) = (a_1 - a_0)\,|\Gamma|\,,$$

where the perimeter function $\mathrm{Per}(\Sigma) := \int_\Omega |\nabla \chi_\Sigma|$ and Γ is the set over which

u is discontinuous. So for binary functions, total variation regularisation is equivalent to both perimeter regularisation and the interfacial length regularisation in the Mumford-Shah model. In fact (5.1) and (5.3) become equivalent.

Suppose that in addition to $u \in BV(\Omega, \{a_0, a_1\})$ we have salt and pepper noise (see 3. in Section 5.1.2). Then the data is binary and both models reduce to the geometric problem

$$\min_{\Sigma_u \subset \Omega} |\Sigma_u \Delta \Sigma_d| + \sigma(a_1 - a_0)\mathrm{Per}(\Sigma_u).$$

Here Σ_u and Σ_d denote respectively the sets where the unknown u and data g_d take the value a_1, $|\cdot|$ is now the N dimensional Hausdorff measure, and $\Sigma_u \Delta \Sigma_d$ is the symmetric difference between the sets.

5.1.3 Phase field model

We base our model on the Mumford-Shah model, but minimise over the space of functions $BV(\Omega, \{a_0, a_1\})$, and generalise it to include the blurring operator S, which we suppose is known a priori. So we have the following nonconvex model with a parameter σ, which we will shortly relax in a different way to (5.3):

$$\operatorname*{arg\,min}_{u \in BV(\Omega, \{a_0, a_1\})} F(u) := \frac{1}{2}\|Su - g_d\|^2_{L^2(\Omega)} + \sigma \mathrm{Per}(\{u = a_1\}). \tag{5.4}$$

We require $S : L^2(\Omega) \to L^2(\Omega)$ to be continuous, linear, and have the mean preservation property i.e. $S1 = 1$ and hence $Sc = c$ for any constant function c. Continuity is helpful for proving existence of minimisers. Linearity and the mean preservation property allow us to recover a function $\bar{u} : \Omega \to \{a_0, a_1\}$ from data g_d by recovering a function $\bar{u} : \Omega \to \{-1, 1\}$ from a scaled and shifted copy of g_d, so long as a_0 and a_1 are known. We assume this to be the case and will therefore restrict our attention to $a_0 = -1$ and $a_1 = 1$ from now onwards.

Some examples of forms S could take are:

1. *Solution operator of elliptic PDE* - Let $Su := y$, where y solves the elliptic boundary value problem

$$\begin{aligned} -\alpha \Delta y + y &= u \quad \text{in } \Omega \\ \frac{\partial y}{\partial \nu} &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{5.5}$$

For any $u \in L^2(\Omega)$ this equation has a unique weak solution $y \in H^1(\Omega)$ which

satisfies the stability estimate

$$\|y\|_{L^2(\Omega)} = \|Su\|_{L^2(\Omega)} \leq C_s(\alpha)\|u\|_{L^2(\Omega)}, \tag{5.6}$$

where $C_s(\alpha) := \frac{1}{1+\alpha/C_p}$ and C_p is the Poincaré constant. So S has all the required properties. We also observe that evaluating S is well-posed, but inverting S is ill-posed, which motivates the need for our model. This is the operator we use for our numerical results.

2. *Convolution operator* - Let

$$Su := \phi_\alpha * u,$$

where ϕ_α is a suitable probability distribution of 'size' α, for example the Gaussian distribution

$$\phi_\alpha(x) = \frac{1}{\alpha\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\alpha^2}\right)$$

of mean zero and variance α, and $*$ is the convolution operation. Such an operator is used in the barcode problem of [Esedoglu, 2004], [Choksi and Gennip, 2010] and [Choksi et al., 2011].

In both of these examples we have a parameter α which controls the extent of the blurring effect. Large α corresponds to heavy blurring and small α corresponds to light blurring. In our work the value of α is known a priori since we assume complete knowledge of S. However there are applications where we may want to relax this assumption, for example the blind deconvolution barcode problem of [Esedoglu, 2004]. In this application we do not know a-priori the distance of the barcode from the scanner, which means the level of blurring is unknown. This can be dealt with by fixing α to be some reasonable guess, or optimising for α at the same time as u.

We relax the model (5.4) by replacing the perimeter functional by the Ginzburg-Landau functional $G_\varepsilon : L^1(\Omega) \to [0, \infty]$ defined by

$$G_\varepsilon(u) := \begin{cases} \int_\Omega \frac{\varepsilon}{2} |\nabla u|^2 + \frac{1}{\varepsilon}\Psi(u) & u \in H^1(\Omega), \\ \infty & \text{otherwise.} \end{cases}$$

for some suitable $\Psi : \mathbb{R} \to \mathbb{R}$, and then minimising over $H^1(\Omega)$ instead of $BV(\Omega, \{-1, 1\})$. So we consider

$$\operatorname*{arg\,min}_{u \in H^1(\Omega)} F_\varepsilon(u) := \frac{1}{2}\|Su - g_d\|^2_{L^2(\Omega)} + \frac{\sigma}{c(\Psi)}\left(\int_\Omega \frac{\varepsilon}{2} |\nabla u|^2 + \frac{1}{\varepsilon}\Psi(u)\right). \tag{5.7}$$

We will focus on two different forms for the potential Ψ; the smooth double well potential

$$\Psi_1(u) := \frac{1}{4}(1-u^2)^2,$$

and the double obstacle potential

$$\begin{aligned}\Psi_2(u) :=& \frac{1}{2}(1-u^2) + I_{[-1,1]}(u) \\ =& \begin{cases} \frac{1}{2}(1-u^2) & |u| \leq 1, \\ \infty & |u| > 1. \end{cases}\end{aligned}$$

This approach, which is called a phase field approximation, results in a diffuse interface with minimisers no longer just taking the values $\{-1, 1\}$, but values in the interval $[-1, 1]$. It is still a nonconvex problem, but it has the advantage of allowing us to minimise over a smoother space of functions for which there is better developed theory. We are able to justify this approach with the following result.

Theorem 5.1. *Let Ψ be the smooth double well potential Ψ_1. Then $G_\varepsilon(u)$ Γ-converges in $L^1(\Omega)$ as $\varepsilon \to 0$ to*

$$\begin{cases} c(\Psi_1)\mathrm{Per}(\{u=1\}) & u \in BV(\Omega, \{-1,1\}), \\ \infty & \textit{otherwise}, \end{cases}$$

where $c(\Psi_1) = 2\int_{-1}^{1} \sqrt{2\Psi_1(s)}\mathrm{d}s = \frac{4\sqrt{2}}{3}$.

Proof. See [Modica and Mortola, 1977]. □

A similar result holds for the double obstacle potential, and performing a calculation we get that $c(\Psi_2) = \frac{\pi}{2}$ (see [Blowey and Elliott, 1993]). To simplify notation we let $\sigma_i = \sigma / c(\Psi_i)$. This ensures that the weighting given to the regularisation is asymptotically σ for both potentials.

The different potentials lead to different formulations and we need to use different approaches to solve them. In particular, Ψ_1 leads to nonlinearity in the zeroth order terms, where as Ψ_2 causes nonlinearity by imposing constraints on the solution.

5.1.4 Literature review

We now mention other parts of the literature which overlap with aspects of this work.

Barcode problem. The 1D version of our problem is related to the barcode problem of Esedoglu in [Esedoglu, 2004]. This work was later extended by Choksi and Gennip in [Choksi and Gennip, 2010]. [Choksi et al., 2011] uses similar ideas on QR barcodes. References for more general image processing literature can be found in Section 5.1.1.

PDE constrained inverse problems. A survey of the literature from the optimal control perspective can be found in [Petra and Stadler, 2011]. In addition, [Tai and Chan, 2004] describes a number of applications where we want to recover piecewise constant functions, such as magnetic resonance imaging (MRI). The thesis [Hackl, 2006] discusses a wide range of techniques for geometric inverse problems. [Tai and Li, 2007] recovers a piecewise constant diffusion coefficient from an elliptic PDE in 2D using the level set method. Recovering an interface from boundary measurements with a different problem formulation is considered in [Kunisch and Pan, 1994]. A related inverse problem that involves recovering piecewise constant functions and uses total variation can be found in [Canelas et al., 2014]

Phase field. In [Hackl, 2006] there is a brief discussion of using a phase field approximation with the smooth double well potential for binary recovery. [Esedoglu, 2004] and [Choksi and Gennip, 2010] use this idea for numerical simulations, though they do not justify the approach analytically. Theory for the phase field approximation with the double obstacle potential can be found in papers by Blowey and Elliott, including [Blowey and Elliott, 1992] and [Blowey and Elliott, 1991]. In [Sarbu, 2010] the double obstacle potential is used in the context of image processing, but without deblurring.

Level set method. This is an alternative way of recovering the discontinuities in our problem. It is discussed in [Tai and Chan, 2004] and [Tai and Li, 2007].

Approximation of Mumford-Shah. [Chambolle and Maso, 1999] and related papers prove Γ-convergence results for finite element approximations of the Mumford-Shah functional. These results have some relation to the convergence results that we obtain using a different approach.

Our work differs from existing work, and hence offers a new contribution, in the following respects:

- We introduce the phase field approximation to the model right from the start (rather than at the last minute in order to allow numerical simulations). We therefore prove rigorous analytical results for this approximate model, which puts our approach on a much firmer footing than in existing work.

- Not only the smooth double well potential, but also the double obstacle potential is used for the phase field approximation. Results are proved for both simultaneously using an abstract framework.
- We thoroughly investigate the dependency of the model on the parameters and perform a systematic comparison of the smooth double well potential and the double obstacle potential on a 1D problem. This highlights some advantages and attractive features of the latter in this setting.

5.1.5 Layout

In Section 5.2 we introduce an abstract optimisation problem, an iterative method for finding critical points of this problem, and prove a convergence result for the iterative method. In Section 5.3 we show that (5.7) fits into this framework with both the smooth double well and double obstacle potentials. In Section 5.4 we discuss a gradient flow formulation of (5.7) and its link to the iterative method. In Section 5.5 we discretise the iterative method and prove another convergence result. We also look at a finite element discretisation for a particular choice of S. In Section 5.6 we demonstrate that implementations of the iterative method work well in 1 and 2 dimensions. In Section 5.7 the performance of using both potentials is compared in detail for a 1D problem. In Section 5.8 we apply our algorithms to a real problem from materials science. In Appendix 5.A we describe how we choose the parameters in our model for the numerical results.

5.2 Abstract framework

Rather than developing separate theory for solving (5.7) with the smooth double well and obstacle potentials, it is advantageous to introduce an abstract framework that both problems fit into.

To this end let V and H be real Hilbert spaces with V compactly embedded in H, and let W be a closed convex nonempty subset of V. Let $b : V \times V \to \mathbb{R}$ and $c : H \times H \to \mathbb{R}$ be symmetric continuous bilinear forms with the properties

$$\begin{aligned} \exists\, \beta \text{ s.t. } b(\eta, \eta) \geq \beta \|\eta\|_V^2 \quad \forall \eta \in V \\ c(\eta, \eta) \geq 0 \quad \forall \eta \in H. \end{aligned}$$

Let $l : V \to \mathbb{R}$ be a bounded linear functional and $J : V \to \mathbb{R}$ a continuous convex

functional. With these objects we can define the energy functional $I : V \to \mathbb{R}$ by

$$I(\eta) := \frac{1}{2}b(\eta, \eta) + J(\eta) - \frac{1}{2}c(\eta, \eta) - l(\eta),$$

which for positive constants α_0 and C_0 we assume satisfies

$$I(\eta) \geq \alpha_0 \|\eta\|_V^2 - C_0 \quad \forall \eta \in W. \tag{5.8}$$

Remark 5.2. *The functional I can be decomposed in different ways into b, J, c and l.*

Optimisation formulation

Consider the following optimisation problem: Find $u \in W$ such that

$$I(u) = \inf_{\eta \in W} I(\eta). \tag{5.9}$$

We can show existence of a solution to (5.9) with the following general result.

Proposition 5.3. *Let $A_1(\cdot) : V \to \mathbb{R}$ be continuous and convex (i.e. weakly lower semicontinuous) and let $A_2(\cdot) : H \to \mathbb{R}$ be continuous. If there exist positive constants α and C such that*

$$A(\eta) := A_1(\eta) + A_2(\eta) \geq \alpha \|\eta\|_V^2 - C \quad \forall v \in V$$

then the following optimisation problem has a solution: Find $u \in W$ such that

$$A(u) = \inf_{\eta \in W} A(\eta).$$

Proof. This follows using the same argument as in Theorem 2.5, which can be found in [Tröltzsch, 2010]. □

Corollary 5.4. *(5.9) has a solution.*

Proof. Take $A_1(\eta) := \frac{1}{2}b(\eta, \eta) + J(\eta) - l(\eta)$ and $A_2(\eta) := -\frac{1}{2}c(\eta, \eta)$. Recall that continuous convex functionals are weakly lower semicontinuous, so A_1 and A_2 satisfy the requirements of Theorem 5.3. □

Note that in general there is not a unique solution to (5.9).

Variational inequality formulation

Solutions to (5.9) must satisfy the following: Find $u \in W$ such that

$$b(u, \eta - u) + J(\eta) - J(u) \geq c(u, \eta - u) + l(\eta - u) \quad \forall \eta \in W. \tag{5.10}$$

Here we have used that J is a convex function, so it has a subdifferential ∂J, which by definition satisfies

$$J(\eta) - J(u) \geq \langle v, \eta - u \rangle \quad \forall v \in \partial J(u),$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between V^* and V. If J is in addition Gâteaux differentiable then (5.10) is equivalent to the following variational inequality: Find $u \in W$ such that

$$b(u, \eta - u) + \langle J'(u), \eta - u \rangle \geq c(u, \eta - u) + l(\eta - u) \quad \forall \eta \in W. \tag{5.11}$$

We often call solutions of (5.10) critical points of (5.9).

Remark 5.5. *If $c(\eta, \eta) \leq \kappa b(\eta, \eta)$ for all $\eta \in V$ with $\kappa < 1$, then (5.10) has a unique solution. When we fit (5.7) into this framework, we find that this would require ε to be large. We intend to take ε small so that (5.7) approximates (5.4), which means we will not necessarily have uniqueness.*

Note that solutions of (5.9) solve (5.10), but the converse is not necessarily true. We nevertheless aim to solve (5.10), as this is much easier in practice. Once a solution has been found, additional tests would have to be used to verify that the solution is a local minimiser of I.

Iterative method

We apply to (5.10) the following generalisation of the iterative method of Barrett and Elliott [Barrett and Elliott, 1991]: Given $u^0 \in W$, for $n = 1, 2, ...$ find $u^n \in W$ such that

$$b(u^n, \eta - u^n) + J(\eta) - J(u^n) \geq c(u^{n-1}, \eta - u^n) + l(\eta - u^n) \quad \forall \eta \in W. \tag{5.12}$$

If J is in addition Gâteaux differentiable then this is equivalent to the following iterative method: Given $u^0 \in W$, for $n = 1, 2, ...$ find $u^n \in W$ such that

$$b(u^n, \eta - u^n) + \langle J'(u^n), \eta - u^n \rangle \geq c(u^{n-1}, \eta - u^n) + l(\eta - u^n) \quad \forall \eta \in W. \tag{5.13}$$

Note that $b(\eta,\eta) + J(\eta)$ is convex and $-c(\eta,\eta) - l(\eta)$ is concave.

(5.12) and (5.13) have unique solutions as they are equivalent to minimising a convex functional over W. Moreover we can prove the following convergence result.

Theorem 5.6. *Every sequence $\{u^n\}$ generated by (5.12) satisfies*

$$I(u^n) + c(u^n - u^{n-1}, u^n - u^{n-1}) + \beta\|u^n - u^{n-1}\|_V^2 \leq I(u^{n-1}) \tag{5.14}$$

and has a subsequence which converges in V to a critical point of (5.9) i.e. a solution of (5.10). Also, the limit of any subsequence of $\{u^n\}$ that converges weakly in V, and hence strongly in H, is a critical point of (5.9).

Proof. The proof is an extension to that of Theorem 6.1 in [Barrett and Elliott, 1991], which proves the same result for a formulation without the J term. To deduce (5.14) we test (5.12) with $\eta = u^{n-1}$ and use the coercivity of b. Because of the assumptions on I, $\{u^n\}$ is uniformly bounded in V, so we can extract a subsequence, which we also denote by $\{u^n\}$, that converges weakly in V and strongly in H to some element $u \in W$. The assumptions on b, c, l and J allow us to pass to the limit in (5.12) and deduce that u satisfies (5.10). The same argument applies to any subsequence, which proves the second part of the theorem.

To see why the convergence in the first part of the theorem is strong in V, note that now we know u satisfies (5.10), we can combine this inequality with (5.12) to get

$$b(u - u^n, u - u^n) \leq c(u - u^{n-1}, u - u^n).$$

The result then follows using the coercivity of b and the strong convergence of u^n in H. □

5.3 Binary recovery application

We now show that (5.7) with both the smooth double well and double obstacle potentials can be fitted into the framework of the previous section.

Smooth double well potential

Set $V, W := H^1(\Omega), H := L^2(\Omega)$, let $S : H \to H$ satisfy the assumptions in Section 5.1.3, and take

$$\begin{aligned} b(u,\eta) &:= (Su, S\eta) + \sigma_1\varepsilon(\nabla u, \nabla\eta) \\ c(u,\eta) &:= \frac{\sigma_1}{\varepsilon}(u,\eta) \\ l(u) &:= (S^* g_d, u) \\ J(u) &:= \frac{\sigma_1}{4\varepsilon}\int_\Omega u^4. \end{aligned}$$

Here and throughout this chapter $(\cdot,\cdot)$ denotes the $L^2(\Omega)$ inner product. S^* denotes the adjoint operator of S, which is defined as follows: For real Hilbert spaces U, V the adjoint operator of a continuous linear operator $A : U \to V$ is the unique continuous linear operator $A^* : V \to U$ such that

$$(Au, v)_V = (u, A^* v)_U \quad \forall u \in U, v \in V.$$

This definition is a special case of the adjoint for operators involving Banach spaces (see (2.5) and (3.4)). The above objects have the properties required in Section 5.2. Coercivity of b can be shown using a contradiction argument and that $S0 = 0$. J is well defined and continuous since $H^1(\Omega)$ is continuously embedded in $L^6(\Omega)$ for $\Omega \subset \mathbb{R}^n$ with $n \leq 3$. I satisfies assumption (5.8) since

$$\int_\Omega \frac{u^4}{4} - \frac{u^2}{2} \geq \int_\Omega \frac{u^2}{2} - 1 = \frac{1}{2}\|u\|^2_{L^2(\Omega)} - |\Omega|\,,$$

and so using that $\|Su - g_d\|^2_{L^2(\Omega)} \geq 0$ we get

$$I(u) \geq \frac{\sigma_1\varepsilon}{2}\|\nabla u\|^2_{L^2(\Omega)} + \frac{\sigma_1}{2\varepsilon}\|u\|^2_{L^2(\Omega)} - \frac{\sigma_1}{\varepsilon}|\Omega| \geq \frac{\sigma_1}{2}\min\left(\varepsilon, \frac{1}{\varepsilon}\right)\|u\|^2_V - \frac{\sigma_1}{\varepsilon}|\Omega|$$

for all $u \in W$. Moreover I equals F_ε from (5.7) with the smooth double well potential (up to an additive constant), so (5.9) becomes: Given $g_d \in L^2(\Omega)$ find

$$\operatorname*{arg\,min}_{u \in H^1(\Omega)} F_1(u) := \frac{1}{2}\|Su - g_d\|^2_{L^2(\Omega)} + \sigma_1\left(\int_\Omega \frac{\varepsilon}{2}|\nabla u|^2 + \frac{1}{\varepsilon}\Psi_1(u)\right). \tag{5.15}$$

J is Gâteaux differentiable, so solutions to (5.15) satisfy (5.11), which be-

comes: Given $g_d \in L^2(\Omega)$, find $u \in H^1(\Omega)$ such that

$$(S^*(Su - g_d), \eta) + \sigma_1 \varepsilon(\nabla u, \nabla \eta) + \frac{\sigma_1}{\varepsilon}(u^3 - u, \eta) = 0 \quad \forall \eta \in H^1(\Omega).$$

In this example we have an equality instead of a variational inequality because W is the full space V.

(5.13) gives the following iterative method for solving the above variational inequality, and it converges by Theorem 5.6: Given $g_d \in L^2(\Omega)$ and $u^0 \in H^1(\Omega)$, for $n = 1, 2, ...$ find $u = u^n \in H^1(\Omega)$ such that

$$(S^*(Su - g_d), \eta) + \sigma_1 \varepsilon(\nabla u, \nabla \eta) + \frac{\sigma_1}{\varepsilon}(u^3 - u^{n-1}, \eta) = 0 \quad \forall \eta \in H^1(\Omega). \tag{5.16}$$

Double obstacle potential

Define $K := \{u \in H^1(\Omega) : |u| \leq 1 \text{ a.e. in } \Omega\}$. Set $V := H^1(\Omega), W := K, H := L^2(\Omega)$, let $S : H \to H$ satisfy the assumptions in Section 5.1.3, and take

$$\begin{aligned} b(u, \eta) &:= (Su, S\eta) + \sigma_2 \varepsilon(\nabla u, \nabla \eta) \\ c(u, \eta) &:= \frac{\sigma_2}{\varepsilon}(u, \eta) \\ l(u) &:= (S^* g_d, u) \\ J(u) &:= 0. \end{aligned}$$

The above objects have the properties required in Section 5.2. As with the smooth double well potential, I satisfies assumption (5.8) since for $u \in W$ we have

$$-\int_\Omega \frac{u^2}{2} \geq \int_\Omega \frac{u^2}{2} - 1 = \frac{1}{2}\|u\|^2_{L^2(\Omega)} - |\Omega|.$$

Moreover I equals F_ε from (5.7) with the double obstacle potential (up to an additive constant), so (5.9) becomes: Given $g_d \in L^2(\Omega)$ find

$$\arg\min_{u \in K} F_2(u) := \frac{1}{2}\|Su - g_d\|^2_{L^2(\Omega)} + \sigma_2 \left(\int_\Omega \frac{\varepsilon}{2} |\nabla u|^2 + \frac{1}{2\varepsilon}(1 - u^2) \right). \tag{5.17}$$

Solutions to (5.17) satisfy (5.11), which becomes: Given $g_d \in L^2(\Omega)$, find $u \in K$ such that

$$(S^*(Su - g_d), \eta - u) + \sigma_2 \varepsilon(\nabla u, \nabla \eta - \nabla u) - \frac{\sigma_2}{\varepsilon}(u, \eta - u) \geq 0 \quad \forall \eta \in K.$$

(5.13) gives the following iterative method for solving the above variational

inequality, which converges by Theorem 5.6: Given $g_d \in L^2(\Omega)$ and $u^0 \in K$, for $n = 1, 2, ...$ find $u = u^n \in K$ such that

$$(S^*(Su - g_d), \eta - u) + \sigma_2\varepsilon(\nabla u, \nabla\eta - \nabla u) - \frac{\sigma_2}{\varepsilon}(u^{n-1}, \eta - u) \geq 0 \quad \forall \eta \in K. \quad (5.18)$$

5.3.1 Alternative iterative methods

In (5.16) and (5.18) the S^*S term is taken implicitly (i.e. we solve for it rather than evaluate it), so we need to be able to invert the operator $S^*S - \sigma_i\varepsilon\Delta$ efficiently, otherwise these iterative methods will be too computationally expensive. In some cases this may be possible, for example if S is the identity, but in general this is not the case.

As we remarked earlier, the definitions of b and c that make I correspond to (5.15) and (5.17) are not unique. For example we can set $b(u, \eta) = B(u, \eta) + \rho(u, \eta)$ and $c(u, \eta) = C(u, \eta) + \rho(u, \eta)$ for some $\rho \geq 0$. The $\rho(u, \eta)$ terms cancel out in I, so defining B and C the same way b and c were defined earlier in this section gives the same optimisation problems (5.15) and (5.17). But the corresponding iterative methods are different. The point of this is that the $\rho(u, \eta)$ term is convex (when $\eta = u$), so it gives us more flexibility in how we define B and C while still having b and c satisfy the coercivity and positivity assumptions.

In particular, for suitably large ρ we can take the S^*S term explicitly (which in our framework corresponds to moving it from b to c), and also take the $\frac{\sigma_i}{\varepsilon}(u, \eta)$ term implicitly (i.e. move it from c to b). So for our examples this corresponds to taking

$$\begin{aligned} b(u, \eta) &:= \rho(u, \eta) + \sigma_i\varepsilon(\nabla u, \nabla\eta) - \frac{\sigma_i}{\varepsilon}(u, \eta), \\ c(u, \eta) &:= \rho(u, \eta) - (S^*Su, \eta). \end{aligned}$$

A restriction such as $\rho > \max(\frac{\sigma_i}{\varepsilon}, C_s^2)$, where C_s is the stability constant from (5.6), is then sufficient for both b to be coercive and c to be nonnegative. So we have the following iterative methods, which are in general easier to solve computationally than (5.16) and (5.18).

Example 5.7 (Smooth double well)**.** *Given $g_d \in L^2(\Omega)$ and $u^0 \in H^1(\Omega)$, for $n = 1, 2, ...$ find $u = u^n \in H^1(\Omega)$ such that*

$$\rho(u - u^{n-1}, \eta) + (S^*(Su^{n-1} - g_d), \eta) + \sigma_1\varepsilon(\nabla u, \nabla\eta) + \frac{\sigma_1}{\varepsilon}(u^3 - u, \eta) = 0 \quad (5.19)$$

for all $\eta \in H^1(\Omega)$.

Example 5.8 (Double obstacle)**.** *Given $g_d \in L^2(\Omega)$ and $u^0 \in K$, for $n = 1, 2, ...$ find $u = u^n \in K$ such that*

$$\rho(u - u^{n-1}, \eta - u) + (S^*(Su^{n-1} - g_d), \eta - u) + \sigma_2\varepsilon(\nabla u, \nabla\eta - \nabla u) - \frac{\sigma_2}{\varepsilon}(u, \eta - u) \geq 0 \tag{5.20}$$

for all $\eta \in K$.

When solving Example 5.7 in practice, it is more convenient for us to solve a linear equation. Therefore we linearise the $J'(u)$ term in (5.19) and consider the following iterative method.

Example 5.9 (Smooth double well)**.** *Given $g_d \in L^2(\Omega)$ and $u^0 \in H^1(\Omega)$, for $n = 1, 2, ...$ find $u = u^n \in H^1(\Omega)$ such that*

$$\rho(u - u^{n-1}, \eta) + (S^*(Su^{n-1} - g_d), \eta) + \sigma_1\varepsilon(\nabla u, \nabla\eta) + \frac{\sigma_1}{\varepsilon}((u^{n-1})^2 u - u, \eta) = 0 \tag{5.21}$$

for all $\eta \in H^1(\Omega)$.

This iterative method lies outside of our framework, so the convergence theory does not necessarily hold. However it works well in practice.

To finish this section we show how we can reformulate the iterative methods to remove $S^*(Su^{n-1} - g_d)$ when S is defined as in (5.5). For example, (5.20) becomes the following.

Example 5.10 (Double obstacle)**.** *Given $g_d \in L^2(\Omega)$ and $u^0 \in K$, for $n = 1, 2, ...$ find $u = u^n \in K$ such that*

$$\rho(u - u^{n-1}, \eta - u) + (p^{n-1}, \eta - u) + \sigma_2\varepsilon(\nabla u, \nabla\eta - \nabla u) - \frac{\sigma_2}{\varepsilon}(u, \eta - u) \geq 0$$

for all $\eta \in K$, where $p^{n-1} \in H^1(\Omega)$ solves

$$\alpha(\nabla p^{n-1}, \nabla\eta) + (p^{n-1}, \eta) = (y^{n-1} - g_d, \eta) \quad \forall \eta \in H^1(\Omega),$$

and y^{n-1} solves the weak form of (5.5) with $u = u^{n-1}$.

5.4 Gradient flow

In this section we investigate the gradient flow method for finding critical points of (5.15) and (5.17) from an initial guess u_0. We prove that this method has some desirable properties, and note the link to the iterative method of the previous sections.

Smooth double well potential

Let u_0 denote the initial guess of the solution and consider the L^2 gradient flow of F_1 in (5.15).

Problem 5.11. *Given $g_d \in L^2(\Omega)$ and $u_0 \in H^1(\Omega)$, find $u \in L^2(0,T;H^1(\Omega))$ with weak time derivative $\partial_t u \in L^2(0,T;L^2(\Omega))$ such that $u(0) = u_0$ and*

$$(\partial_t u(t), \eta) + (S^*(Su(t) - g_d), \eta) + \sigma_1 \varepsilon (\nabla u(t), \nabla \eta) + \frac{\sigma_1}{\varepsilon}(\Psi_1'(u(t)), \eta) = 0 \quad (5.22)$$

for all $\eta \in H^1(\Omega)$ and almost all $t \in (0,T)$.

Theorem 5.12. *Problem 5.11 has a unique solution.*

Proof. Note that Problem 5.11 is very similar to the Allen-Cahn equation with the smooth double well potential, and the proof follows using standard techniques. See for example the references in Theorem 5.15, where existence and uniqueness is proved for smooth potentials in order to show existence and uniqueness for the double obstacle potential in the limit. □

Theorem 5.13. *If u is a sufficiently smooth solution of Problem 5.11 then the energy $F_1(u(t))$ decreases over time.*

Proof. For some $t \in (0,T)$ we can test (5.22) with $\eta = \partial_t u(t)$ to get

$$\begin{aligned}\|\partial_t u(t)\|_{L^2(\Omega)}^2 + (S^*(Su(t) - g_d), \partial_t u(t)) + \sigma_1 \varepsilon (\nabla u(t), \nabla \partial_t u(t)) \\ + \frac{\sigma_1}{\varepsilon}(\Psi_1'(u(t)), \partial_t u(t)) = 0. \quad (5.23)\end{aligned}$$

Note that

$$\begin{aligned}(S^*(Su(t) - g_d), \partial_t u(t)) &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|Su(t) - g_d\|_{L^2(\Omega)}^2, \\ (\nabla u(t), \nabla \partial_t u(t)) &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\nabla u(t)\|_{L^2(\Omega)}^2, \\ (\Psi_1'(u(t)), \partial_t u(t)) &= \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \Psi_1(u(t)),\end{aligned}$$

so equation (5.23) is equivalent to

$$\|\partial_t u(t)\|_{L^2(\Omega)}^2 + \frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{1}{2}\|Su(t) - g_d\|_{L^2(\Omega)}^2 + \frac{\sigma_1\varepsilon}{2}\|\nabla u(t)\|_{L^2(\Omega)}^2 + \frac{\sigma_1}{\varepsilon}\int_\Omega \Psi_1(u(t))\Big) = 0.$$

Therefore as long as $\partial_t u(t)$ is not zero almost everywhere we have

$$0 > -\|\partial_t u(t)\|_{L^2(\Omega)}^2 \geq \frac{\mathrm{d}}{\mathrm{d}t} F_1(u(t)),$$

and hence the energy decreases. □

Double obstacle potential

We can formulate a gradient flow for F_2 from (5.17) in a similar way.

Problem 5.14. *Given $g_d \in L^2(\Omega)$ and $u_0 \in H^1(\Omega)$, find $u \in K_T$ with $\partial_t u \in L^2(0,T;L^2(\Omega))$ such that $u(0) = u_0$ and*

$$\begin{aligned}(\partial_t u(t), \eta - u(t)) + (S^*(Su(t) - g_d), \eta - u(t)) + \sigma_2\varepsilon(\nabla u(t), \nabla\eta - \nabla u(t)) \\ - \frac{\sigma_2}{\varepsilon}(u(t), \eta - u(t)) \geq 0\end{aligned} \tag{5.24}$$

for all $\eta \in K$ and almost all $t \in (0,T)$. Here

$$K_T := \{u \in L^2(0,T;H^1(\Omega)) : |u| \leq 1 \text{ a.e. in } (0,T)\times\Omega\}.$$

Theorem 5.15. *Problem 5.14 has a unique solution. Moreover, if u is a sufficiently smooth solution then the energy $F_2(u(t))$ decreases over time.*

Proof. This follows from a slight modification to the arguments for the double obstacle Allen-Cahn inequality in [Blank et al., 2011; Chen and Elliott, 1994; Blank et al., 2012; Blowey and Elliott, 1993, 1991] to allow for the S^*Su term. □

For both potentials it is important to consider whether $u(t)$ converges to a steady state as $t \to \infty$, and whether this steady state is a critical point. These types of issues are investigated in [Hale, 1988], and in [Chen and Elliott, 1994] for the 1D double obstacle potential. We do not discuss this as the focus of this work is on iterative methods.

5.4.1 Link to iterative methods

Particular first order discretisations in time of the gradient flow formulations are equivalent to the iterative methods of the previous section with $\rho = \frac{1}{\Delta t}$. But we only want to solve the optimisation problems (5.15) and (5.17); we are not interested in the accuracy of solutions to (5.22) and (5.24) at each point in time, but rather how well they approximate minimisers of F_1 and F_2 for large t. For this reason our method for solving (5.15) and (5.17) should focus on decreasing the energy. The iterative methods of the previous sections are designed to have this property, where as discretisations in time of the gradient flows may not.

The scheme denoted by (5.12) of Barrett and Elliott motivated the convexity splitting implicit/explicit Euler scheme used in [Elliott and Stuart, 1993]. See also [Eyre, 1998].

5.5 Discretisation

In this section we discretise the abstract iterative method of Section 5.2 in space and analyse convergence of the discretisation. We then apply this theory to a finite element discretisation of (5.19) and (5.20) for S defined by 5.5.

5.5.1 Discrete abstract framework

Suppose we have a family of subspaces $V_h \subset V$ and closed convex nonempty subsets $W_h \subset V_h$ which approximate functions in W increasingly well as some parameter $h \to 0$. In particular we suppose we have an approximation operator $P_h : W \to W_h$ such that

$$\|\eta - P_h\eta\|_V \to 0 \text{ as } h \to 0 \quad \forall \eta \in W, \tag{5.25}$$

and that every sequence $\{\eta_h\} \subset W_h$ satisfies

$$\eta_h \rightharpoonup \eta \text{ in } V \text{ as } h \to 0 \implies \eta \in W. \tag{5.26}$$

Remark 5.16. *Note that we do not require $W_h \subset W$. If this holds then (5.26) follows automatically because W is a closed convex subset of a Banach space, and hence is weakly sequentially closed.*

We now assume there exist objects b_h, c_h and l_h which satisfy the same assumptions as b, c and l, with the boundedness and coercivity constants independent of h. We define

$$I_h(\eta) := \frac{1}{2}b_h(\eta,\eta) + J(\eta) - \frac{1}{2}c_h(\eta,\eta) - l_h(\eta),$$

and as in (5.8) we assume that there exist positive constants α_1 and C_1 independent of h such that

$$I_h(\eta_h) \geq \alpha_1\|\eta_h\|_V^2 - C_1 \quad \forall \eta_h \in W_h. \tag{5.27}$$

So minimisers of I_h over W_h (which exist, since I_h satisfies the same assumptions

as I) satisfy the following discrete problem: Find $u_h \in W_h$ such that

$$b_h(u_h, \eta_h - u_h) + J(\eta_h) - J(u_h) \geq c_h(u_h, \eta_h - u_h) + l_h(\eta_h - u_h) \quad \forall \eta_h \in W_h. \tag{5.28}$$

If J is in addition Gâteaux differentiable then this is equivalent to the following discrete variational inequality: Find $u_h \in W_h$ such that

$$b_h(u_h, \eta_h - u_h) + \langle J'(u_h), \eta_h - u_h \rangle \geq c_h(u_h, \eta_h - u_h) + l_h(\eta_h - u_h) \quad \forall \eta_h \in W_h.$$

We need b_h, c_h and l_h to approximate their continuous counterparts as $h \to 0$. So we make the additional assumptions that for any bounded sequence $\{v_h\} \subset W$ we have

$$\begin{aligned} \|(b - b_h)(v_h, \cdot)\|_{V^*} &= \sup_{\eta_h \in V_h \setminus \{0\}} \frac{|b(v_h, \eta_h) - b_h(v_h, \eta_h)|}{\|\eta_h\|_V} \to 0, \\ \|(c - c_h)(v_h, \cdot)\|_{V^*} &= \sup_{\eta_h \in V_h \setminus \{0\}} \frac{|c(v_h, \eta_h) - c_h(v_h, \eta_h)|}{\|\eta_h\|_V} \to 0, \\ \|l - l_h\|_{V^*} &= \sup_{\eta_h \in V_h \setminus \{0\}} \frac{|l(\eta_h) - l_h(\eta_h)|}{\|\eta_h\|_V} \to 0 \end{aligned} \tag{5.29}$$

as $h \to 0$. With these assumptions solutions of the discrete variational inequality (5.28) approximate solutions of the continuous variational inequality (5.10) as $h \to 0$, as the following theorem shows.

Theorem 5.17. *For any sequence $h_n \to 0$ the sequence $\{u_{h_n}\}$ of solutions to (5.28) has a subsequence which converges weakly in V, and hence strongly in H, to a critical point of (5.9) i.e. a solution of (5.10). Moreover, the limit of any subsequence of $\{u_{h_n}\}$ that converges weakly in V, and hence strongly in H, is a critical point of (5.9).*

Proof. For a given h we can find $u_h = \arg\min_{\eta_h \in W_h} I_h(\eta_h)$, then for any $\eta_h \in W_h$,

$$I_h(u_h) \leq I_h(\eta_h) = \frac{1}{2} b_h(\eta_h, \eta_h) + J(\eta_h) - \frac{1}{2} c_h(\eta_h, \eta_h) - l_h(\eta_h).$$

Fix $\eta \in W$ and set $\eta_h = P_h \eta \in W_h$. So $\{\eta_h\}$ is bounded in V by (5.25), which means $|b_h(\eta_h, \eta_h) - b(\eta_h, \eta_h)| \leq \|(b_h - b)(\eta_h, \cdot)\|_{V^*} \|\eta_h\|_V \leq C$. Here and throughout this section C denotes a generic constant independent of h which may vary from line to line. A similar result holds for l_h, and c_h is nonnegative, so

$$I_h(u_h) \leq \frac{1}{2} b(\eta_h, \eta_h) + J(\eta_h) + |l(\eta_h)| + C.$$

By the boundedness of b and l,

$$I_h(u_h) \le C(\|\eta_h\|_V^2 + J(\eta_h) + \|\eta_h\|_V).$$

Combining this with (5.27) we get

$$\|u_h\|_V \le C(\|\eta_h\|_V + J(\eta_h) + 1).$$

Now (5.25) and the continuity of J give that $J(\eta_h) \le C$. In addition (5.25) implies that for h less than some h_0, $\|\eta_h\|_V \le \|\eta\|_V + C$, and therefore $\|u_h\|_V \le C$.

From the above it follows that for any sequence $h_n \to 0$, $\{u_{h_n}\}$ is bounded in V. So we can find a subsequence, which we also denote by $\{u_{h_n}\}$, that converges weakly in V and strongly in H to some $u \in V$. In fact $u \in W$ by (5.26). We now show that u is a solution of (5.10).

Note that for all $\eta \in W$ we have

$$\begin{aligned}
&\liminf_{n\to\infty} b_{h_n}(u_{h_n}, P_{h_n}\eta - u_{h_n}) \\
&= \liminf_{n\to\infty} \Big(b_{h_n}(u_{h_n}, P_{h_n}\eta - u_{h_n}) \pm b(u_{h_n}, P_{h_n}\eta - u_{h_n}) \pm b(u_{h_n}, \eta - u_{h_n}) \Big) \\
&= \liminf_{n\to\infty} \Big((b_{h_n} - b)(u_{h_n}, P_{h_n}\eta - u_{h_n}) + b(u_{h_n}, P_{h_n}\eta - \eta) + b(u_{h_n}, \eta - u_{h_n}) \Big) \\
&= \liminf_{n\to\infty} b(u_{h_n}, \eta - u_{h_n}) \\
&\le b(u, \eta - u).
\end{aligned}$$

The final equality follows because $\lim_{n\to\infty}(b_{h_n} - b)(u_{h_n}, P_{h_n}\eta - u_{h_n}) = 0$ by (5.29) and $\lim_{n\to\infty} b(u_{h_n}, P_{h_n}\eta - \eta) = 0$ by (5.25). The inequality follows from the lower semicontinuity of $b(\cdot,\cdot)$ and the continuity of $b(\cdot,\eta)$. We can also show that

$$\liminf_{n\to\infty} c_{h_n}(u_{h_n}, P_{h_n}\eta - u_{h_n}) \ge c(u, \eta - u)$$

using weak lower semicontinuity, and a similar result holds for l_h. This and the continuity and weak lower semicontinuity of J gives

$$\begin{aligned}
b(u, \eta - u) + J(\eta) - J(u) &\ge \liminf_{n\to\infty} \Big(b_{h_n}(u_{h_n}, P_{h_n}\eta - u_{h_n}) + J(P_{h_n}\eta) - J(u_{h_n}) \Big) \\
&\ge \liminf_{n\to\infty} \Big(c_{h_n}(u_{h_n}, P_{h_n}\eta - u_{h_n}) + l_{h_n}(P_{h_n}\eta - u_{h_n}) \Big) \\
&\ge c(u, \eta - u) + l(\eta - u) \quad \forall \eta \in W.
\end{aligned}$$

Hence u is indeed a solution of (5.10).

The same argument applies to any weakly convergent subsequence, which proves the second part of the theorem. □

Remark 5.18. *We could also assume we have functionals J_h satisfying the same assumptions as J, with the continuity independent of h, plus the additional property that $v_{h_n} \rightharpoonup v$ in W for $h_n \to 0$ implies $\liminf_{n\to\infty} J_{h_n}(v_{h_n}) \geq J(v)$. Then a proof almost identical to the above gives convergence for (5.28) with J replaced by J_h. This allows numerical integration to be used on the J term.*

As with (5.10) in Section 5.2, we can consider an iterative method for solving (5.28): Given $u_h^0 \in W_h$, for $n = 1, 2, ...$ find $u_h^n \in W_h$ such that

$$b_h(u_h^n, \eta_h - u_h^n) + J(\eta_h) - J(u_h^n) \geq c_h(u_h^{n-1}, \eta_h - u_h^n) + l_h(\eta_h - u_h^n) \quad \forall \eta_h \in W_h.$$

If J is in addition Gâteaux differentiable then this is equivalent to the following iterative method: Given $u_h^0 \in W_h$, for $n = 1, 2, ...$ find $u_h^n \in W_h$ such that

$$b_h(u_h^n, \eta_h - u_h^n) + (J'(u_h^n), \eta_h - u_h^n) \geq c_h(u_h^{n-1}, \eta_h - u_h^n) + l_h(\eta_h - u_h^n) \quad \forall \eta_h \in W_h.$$

Since b_h, c_h and l_h satisfy the same assumptions as b, c, and l, the above iterative method still has the energy decreasing property, and we get convergence of iterates to a solution of (5.28). Then as $h \to 0$ the solutions of (5.28) converge to critical points of (5.9) by Theorem 5.17.

5.5.2 Finite element discretisation of (5.19) and (5.20)

Assume that Ω is polyhedral and let $\{T_h\}$ be a family of uniform regular triangulations of Ω into disjoint open simplices with a maximal element size h. Associated with each T_h we have the piecewise linear finite element space

$$V_h := \{v \in C^0(\bar{\Omega}) : v|_T \in P_1(T) \text{ for all } T \in T_h\} \subset H^1(\Omega),$$

where $P_1(T)$ is the set of all linear affine functions on T. Also define

$$K_h := \{v_h \in V_h : |v_h| \leq 1 \text{ in } \Omega\}$$

so that we have a finite element space analogous to K. Note that $K_h \subset K$ so Remark 5.16 applies. Take P_h to be the operator that maps $u \in W$ to the unique $P_h u \in W_h$ such that

$$(P_h u, \eta_h - u)_{H^1(\Omega)} \geq (u, \eta_h - u)_{H^1(\Omega)} \quad \forall \eta_h \in W_h.$$

This operator satisfies equation (5.25), see e.g. Chapter 2 in [Glowinski, 1984].

Let S be the solution operator of (5.5), and denote by S_h the discrete blurring operator. We intend this to approximate S, so we define S_h to map $u \in L^2(\Omega)$ to the unique $y_h \in V_h$ satisfying

$$\alpha(\nabla y_h, \nabla \eta_h) + (y_h, \eta_h) = (u, \eta_h) \quad \forall \eta_h \in V_h. \tag{5.30}$$

A stability estimate the same as (5.6) holds, so

$$\|y_h\|_{L^2(\Omega)} = \|S_h u\|_{L^2(\Omega)} \leq C_s(\alpha)\|u\|_{L^2(\Omega)}, \tag{5.31}$$

where as before $C_s(\alpha) = \frac{1}{1+\alpha/C_p}$. Also standard error analysis for elliptic PDEs says

$$\|y - y_h\|_{L^2(\Omega)} \leq Ch\|y\|_{H^1(\Omega)},$$

which combined with (5.31) gives that

$$\|(S - S_h)u\|_{L^2(\Omega)} \leq Ch\|Su\|_{H^1(\Omega)} \leq Ch\|u\|_{L^2(\Omega)}. \tag{5.32}$$

Example 5.19 (Smooth double well)**.** *Take the same definitions as in Example 5.7. In addition take V_h as above, $W_h := V_h$, and define*

$$\begin{aligned} b_h(u_h, \eta_h) &:= \rho(u_h, \eta_h) + \sigma_1\varepsilon(\nabla u_h, \nabla \eta_h) - \frac{\sigma_1}{\varepsilon}(u_h, \eta_h), \\ c_h(u_h, \eta_h) &:= \rho(u_h, \eta_h) - (S_h u_h, S_h \eta_h), \\ l_h(u_h) &:= (S_h^* g_{d,h}, u_h), \\ J(u_h) &:= \frac{\sigma_1}{4\varepsilon}\int_\Omega u_h^4, \end{aligned}$$

where S_h is the discrete elliptic operator defined by (5.30), and $g_{d,h}$ is the L^2-projection of g_d onto V_h.

For $\rho > \max(\frac{\sigma_1}{\varepsilon}, C_s^2)$, where C_s is the stability constant from (5.31), all the assumptions of Theorem 5.6 are satisfied, so we get the decreasing energy property and convergence of iterates for the following discrete iterative method: Given $g_{d,h}$, $u_h^0 \in V_h$, for $n = 1, 2, ...$ find $u_h = u_h^n \in V_h$ such that

$$\begin{aligned} \rho(u_h - u_h^{n-1}, \eta_h) + (p_h^{n-1}, \eta_h) + \sigma_1\varepsilon(\nabla u_h, \nabla \eta_h) \\ + \frac{\sigma_1}{\varepsilon}(u_h^3 - u_h, \eta_h) = 0 \quad \forall \eta_h \in V_h, \end{aligned}$$

where y_h^{n-1}, $p_h^{n-1} \in V_h$ *satisfy*

$$\begin{aligned}\alpha(\nabla y_h^{n-1}, \nabla \eta_h) + (y_h^{n-1}, \eta_h) &= (u_h^{n-1}, \eta_h)\\ \alpha(\nabla p_h^{n-1}, \nabla \eta_h) + (p_h^{n-1}, \eta_h) &= (y_h^{n-1} - g_{d,h}, \eta_h)\end{aligned}$$

for all $\eta_h \in V_h$.

The assumptions of Theorem 5.17 are also satisfied, since for a weakly convergent sequence $\{v_h\} \in V$ *we have*

$$\begin{aligned}|(c_h - c)(v_h, \eta_h)| &= |(S_h v_h, S_h \eta_h) - (S v_h, S \eta_h)|\\ &\leq |(S_h v_h, (S_h - S)\eta_h)| + |((S_h - S)v_h, S\eta_h)|\\ &\leq \|S_h v_h\|_{L^2(\Omega)} \|(S_h - S)\eta_h\|_{L^2(\Omega)} + \|(S_h - S)v_h\|_{L^2(\Omega)} \|S\eta_h\|_{L^2(\Omega)}.\end{aligned}$$

Now using (5.31) and (5.32) we get

$$\|(c_h - c)(v_h, \cdot)\|_{H^1(\Omega)^*} \leq Ch \|v_h\|_{H^1(\Omega)},$$

and so $\|(c_h - c)(v_h, \cdot)\|_{H^1(\Omega)^*} \to 0$ *as* $h \to 0$ *by the boundedness of* $\|v_h\|_V$. *Similar results hold for* b_h *and* l_h. *Therefore we have convergence of limit points of the above discrete iterative method to critical points of (5.15) as* $h \to 0$.

Remark 5.20. *As mentioned before Example 5.9, when solving the smooth double well problem in practice, we solve a finite element discretisation of the linearised iterative method (5.21): Given* $g_{d,h}$, $u_h^0 \in V_h$, *for* $n = 1, 2, ...$ *find* $u_h = u_h^n \in V_h$ *such that*

$$\begin{aligned}\rho(u_h - u_h^{n-1}, \eta_h) + (p_h^{n-1}, \eta_h) + \sigma_1 \varepsilon(\nabla u_h, \nabla \eta_h)\\ + \frac{\sigma_1}{\varepsilon}((u_h^{n-1})^2 u_h - u_h, \eta_h) = 0 \quad \forall \eta_h \in V_h,\end{aligned} \tag{5.33}$$

where y_h^{n-1}, $p_h^{n-1} \in V_h$ *satisfy*

$$\begin{aligned}\alpha(\nabla y_h^{n-1}, \nabla \eta_h) + (y_h^{n-1}, \eta_h) &= (u_h^{n-1}, \eta_h)\\ \alpha(\nabla p_h^{n-1}, \nabla \eta_h) + (p_h^{n-1}, \eta_h) &= (y_h^{n-1} - g_{d,h}, \eta_h)\end{aligned} \tag{5.34}$$

for all $\eta_h \in V_h$.

We use numerical integration on the linearised term. Note that the theorems do not necessarily hold for this iterative method, but it performs well in practice.

Example 5.21 (Double obstacle)**.** *Take the same definitions as in Example 5.8. In*

addition take V_h as above, $W_h := K_h$, and define

$$\begin{aligned} b_h(u_h, \eta_h) &:= \rho(u_h, \eta_h) + \sigma_2 \varepsilon(\nabla u_h, \nabla \eta_h) - \frac{\sigma_2}{\varepsilon}(u_h, \eta_h) \\ c_h(u_h, \eta_h) &:= \rho(u_h, \eta_h) - (S_h u_h, S_h \eta_h) \\ l_h(u_h) &:= (S_h^* g_{d,h}, u_h) \\ J(u_h) &:= 0, \end{aligned}$$

where S_h is the discrete elliptic operator defined by (5.30), and $g_{d,h}$ is the L^2-projection of g_d onto V_h.

For $\rho > \max(\frac{\sigma_2}{\varepsilon}, C_s^2)$ all the assumptions of Theorem 5.6 are satisfied, so we get the decreasing energy property and convergence of iterates for the following discrete iterative method: Given $g_{d,h} \in V_h$ and $u_h^0 \in K_h$, for $n = 1, 2, ...$ find $u_h = u_h^n \in K_h$ such that

$$\begin{aligned} \rho(u_h - u_h^{n-1}, \eta_h - u_h) + (p_h^{n-1}, \eta_h - u_h) + \sigma_2 \varepsilon(\nabla u_h, \nabla \eta_h - \nabla u_h) \\ - \frac{\sigma_2}{\varepsilon}(u_h, \eta_h - u_h) \geq 0 \quad \forall \eta_h \in K_h \end{aligned} \tag{5.35}$$

where y_h^{n-1}, $p_h^{n-1} \in V_h$ satisfy

$$\begin{aligned} \alpha(\nabla y_h^{n-1}, \nabla \eta_h) + (y_h^{n-1}, \eta_h) &= (u_h^{n-1}, \eta_h) \\ \alpha(\nabla p_h^{n-1}, \nabla \eta_h) + (p_h^{n-1}, \eta_h) &= (y_h - g_{d,h}, \eta_h) \end{aligned}$$

for all $\eta_h \in V_h$.

Theorem 5.17 gives convergence of limit points of the above discrete iterative method to critical points of (5.17) as $h \to 0$.

5.5.3 Algorithms

The discrete iterative methods in Examples 5.19 and 5.21 lead to the following algorithms for binary image recovery, which we implement and test in the next section.

Note that despite the blurring and noise, $g_{d,h}$ still contains a lot of information about the solution. Therefore in practice we scale and threshold $g_{d,h}$ in order to get a good initial guess for u_h^0.

Smooth double well potential

Given $g_{d,h} \in V_h$ and an initial guess $u_h^0 \in V_h$, set $n = 1$ then:

1. Solve (5.34) for y_h^{n-1} then p_h^{n-1};
2. Solve (5.33) for u_h^n;
3. If $\|u_h^n - u_h^{n-1}\|_{L^2(\Omega)} < \text{TOL}$ terminate the algorithm. Else set $n = n + 1$ and go to step 1;

An alternative stopping criterion would be to wait until the change in energy $\left|F_1(u_h^n) - F_1(u_h^{n-1})\right|$ is sufficiently small. This has the advantage that the energy decreasing result then guarantees our algorithm terminates. However the stopping criterion in the above algorithm also gives a good indication of a steady state, and we found it easier to calibrate (choose a value that works well for a wide range of problems).

Double obstacle potential

The algorithm for this potential is the same as for the smooth double well potential, but we instead solve the variational inequality (5.35) in step 2.

One method for solving the variational inequalities at each iteration is the primal-dual active set (PDAS) method. It is applied to solving the variational inequalities arising in the Allen-Cahn inequality in [Blank et al., 2012]. We implemented this method and found it to work well. However, for the numerical results in the next sections we use an alternative method known as the Truncated Nonsmooth Newton Multigrid (TNNMG) method (see [Gräser, 2011; Gräser and Kornhuber, 2009]), which performs very well. We are grateful to Carsten Gräser for sharing his Dune-Solvers code for the TNNMG method.

5.6 Numerical results

In this section we show some numerical examples of binary recovery in 1 and 2 dimensions. The data is blurred by the solution operator of the elliptic PDE (5.5), with the parameter α controlling the level of blurring. It also has additive Gaussian noise of mean zero and variance γ.

We do the recovery using the discrete iterative methods of Remark 5.20 (based on the smooth double well potential) and (5.35) (based on the double obstacle potential). In practice we observe convergence of the full sequence of iterates to steady states, which are discrete critical points of (5.7). As we take ε and h small, we believe that these critical points closely approximate a global minimiser of the model (5.4). This is because the iterative methods give us discrete critical points

of the approximate model (5.7), which seem to be at least discrete local minimisers of (5.7), as different initial iterates and (valid) values of ρ do not lead to different steady states. In addition, for small ε (and appropriate h) the critical points are close to being binary i.e. feasible minimisers of the model (5.4). We cannot be certain how close they really are to the global minimisers of (5.4) due to the lack of explicitly known global minimisers for interesting problems. Regardless, by artificially generating data from a known binary function, the numerical results show that for small ε (and h small relative to ε) our iterative methods are effective at recovering something close to the binary function.

The weighting given to the regularisation (the parameter σ), which defines the nonconvex model (5.4), is an important but challenging issue. If we take σ too small then recovered functions still have artifacts of the noise. If σ is too large then we loose some features we actually want to keep. See Figure 5.15 in Appendix 5.A.1 for a comparison. In Appendix 5.A.1 we also discuss some results from the literature on the choice of σ for related problems, however the theory does not apply to our particular problem. In this section we just take values of σ that we have experimentally determined to work well for the problem at hand. Note that more sophisticated and theoretically justified techniques exist. Morozov's discrepancy principle (see [Morozov, 1966]) can be used to estimate σ during the minimisation process. [Osher et al., 2005] and [Wen and Chan, 2012] use this idea on problems with total variation regularisation. A very different statistical approach to solving inverse problem is to use a hierarchical model based on a Gaussian prior for the noise. The posterior distribution for $\bar{u}$ can then be sampled using Markov Chain Monte Carlo. A big advantage of this is that it allows us to understand the uncertainty associated with our estimate for $\bar{u}$. Some literature related to this approach includes [Kaipio and Somersalo, 2005], [Calvetti and Somersalo, 2008] and [Agapiou et al., 2013].

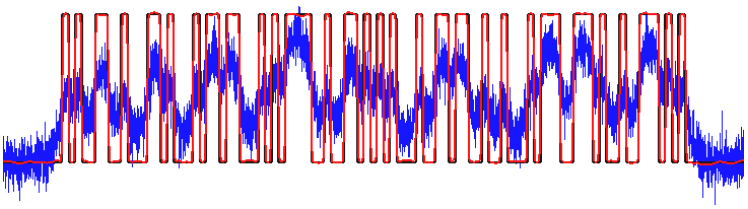
For the implementation we use the Distributed and Unified Numerics Environment (DUNE), see [Blatt and Bastian, 2007; Bastian et al., 2008b,a, 2011; Dedner et al., 2010, 2011]. DUNE provides interfaces for grids, solvers and finite element spaces. Therefore once the algorithms are implemented, it takes minimal effort to change features of the implementation that would usually be fixed, such as the grid type, the dimension of the problem, and the type of finite elements used.

5.6.1 1D numerical results

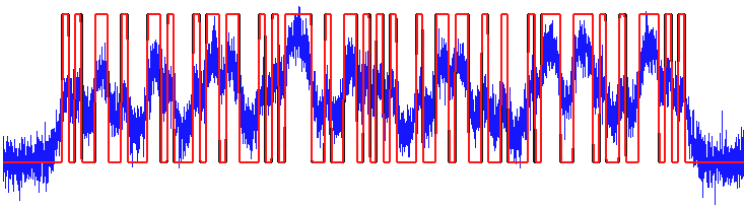
The test problem in 1D is inspired by the barcode problem of [Esedoglu, 2004], which was mentioned as a motivating example in Section 5.1.1. We try to recover a

binary function taking the values $\{-1, 1\}$, which one can imagine represents a cross section of a barcode (with values of -1 corresponding to black parts of the barcode and values of 1 corresponding to white parts). We suppose this binary function is corrupted, giving blurred and noisy data that we want to decode. The main difference between our test problem and the barcode problem in [Esedoglu, 2004] is that we have chosen blurring caused by the solution operator of an elliptic PDE instead of a convolution. Although this is not a realistic blurring operator specified by this application, if our approach is effective for this blurring operator then it is likely to be effective for other blurring operators.

The recovery using both the smooth double well and double obstacle potentials can be found in Figure 5.1. The black lines represent the binary function that we want to recover, the blue lines are the artificial data we generate by adding blurring and noise, and the red lines are the recovered functions for each potential. Even by eye it is not clear exactly how many 'bars' are in the binary functions, or the correct widths of the bars. But the recovered functions closely match the binary function we started with (which is why the black lines are almost hidden by the red lines), showing that our approach is effective. The figure also makes apparent one of the advantages of the double obstacle potential, which is that recovered functions take a form closer to what we actually want; binary functions.



(a) Smooth double well potential.



(b) Double obstacle potential.

Figure 5.1: $\alpha = 1e-4, \gamma = 0.4, \sigma = 1e-4, \varepsilon = 5.31e-4$ and $h = 1.67e-4$.

5.6.2 2D numerical results

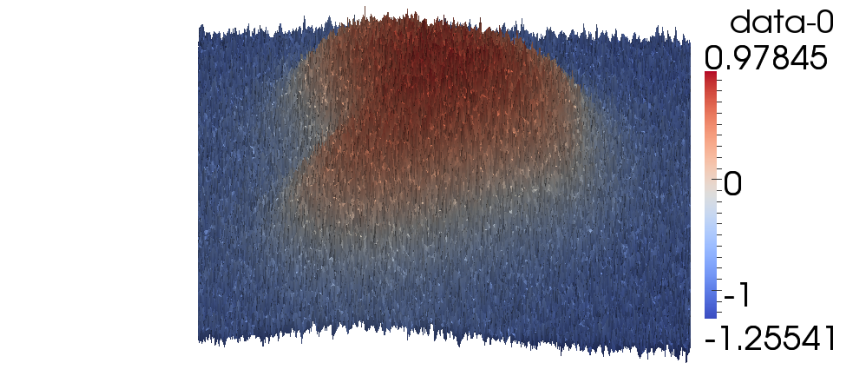
The test problems in 2D involve recovering binary functions with discontinuities of various shapes. In this dimension the problems have a natural interpretation as deblurring and denoising of images, but we also view them as binary source recovery problems for elliptic PDEs.

Figure 5.2 shows the recovery of a binary function using (5.7) with the smooth double well potential. The discontinuity is a 'blob' shape and is marked by a black line. The blurred and noisy data for this function is shown in Figures 5.2(a) and 5.2(b). Figure 5.2(c) shows the recovered function, with a yellow line marking the zero level set. We can see that the yellow line closely matches the black line, except for a slight mismatch at the concave parts of the discontinuity. Note that we cannot make the interface as small as for the 1D problem as the resolution of the grid needed to resolve it makes this computationally expensive. Our implementation is capable of adaptivity, which lessens this cost somewhat, but we will not demonstrate this functionality in this work. With this simple visualisation the recovered function using the double obstacle potential looks very similar, so we do not include a figure of it.

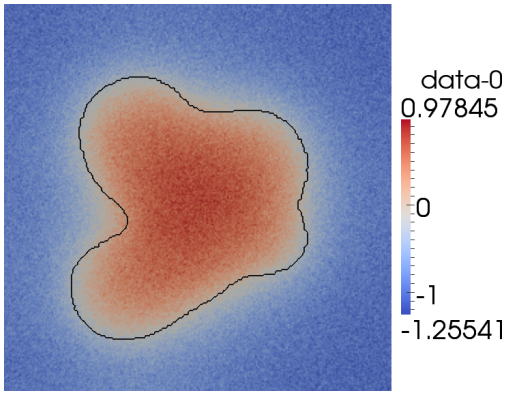
Figure 5.3 (which can be interpreted in the same way as Figure 5.2) shows the recovery of a binary function with a letter 'A' shaped discontinuity. This time we use the double obstacle potential in (5.7), though the recovered function using the smooth double well potential looks similar. This example shows that the model can also recover discontinuities with corners reasonably accurately, but there is some rounding of these corners due to the regularisation.

To finish this section we show an example which relates to an application of binary image recovery in 2D. Figure 5.4 shows the recovery of a binary function representing a QR code with 25x25 blocks (the size typically used to encode a URL). The yellow lines mark the discontinuity of the binary function. Figure 5.4(a) shows the data with a red line marking the zero level set, and Figure 5.4(b) shows the recovered function. We see that features which are blurred below the zero level set (and which therefore would not be recovered by a simple projection) are nevertheless recovered by the model.

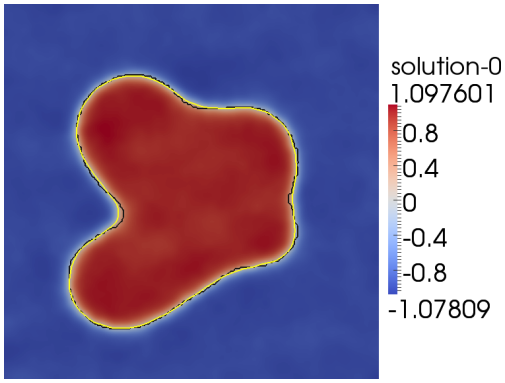
In Section 5.8 we will show further numerical results in 2D, as we apply our algorithms to solve a problem from materials science.



(a)



(b)



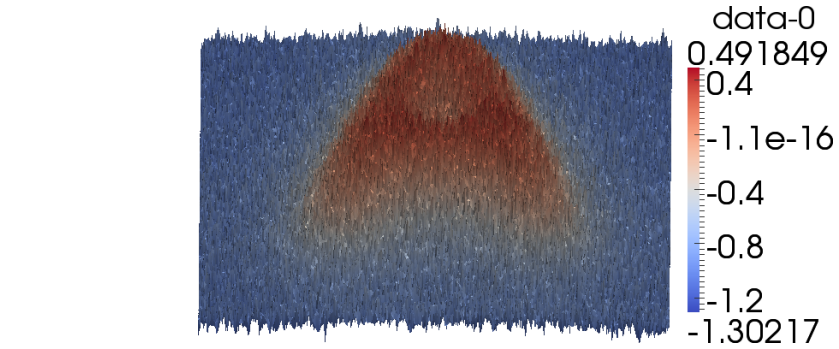
(c)

Figure 5.2: $\alpha = 0.01$, $\gamma = 0.2$, $\sigma = 1e-4$, $\varepsilon = 0.00879$ and $h = 0.00345$ using the smooth double well potential.

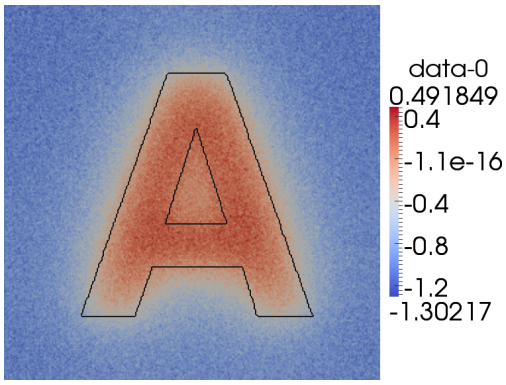
5.7 Comparison of potentials in 1D

Due to the Γ-convergence result of Theorem 5.1, we expect that critical points of (5.7) for a given value of σ using either the smooth double well or double obstacle potential will converge to critical point of (5.4) in the limit of small ε. Of course the critical points they converge to are not guaranteed to be the same, but agreement of the limits is observed in practice, and for very small ε the recovered functions for both potentials are almost indistinguishable. However it is well known that for phase field type problems, the interface should be well resolved in order for an accurate spatial approximation. This means that the smaller ε, the more grid points needed, and the higher the computational cost of the iterative methods. For many applications we only want to recover the location of the discontinuities in a binary function, which we suppose are given by the zero level set of the recovered function.

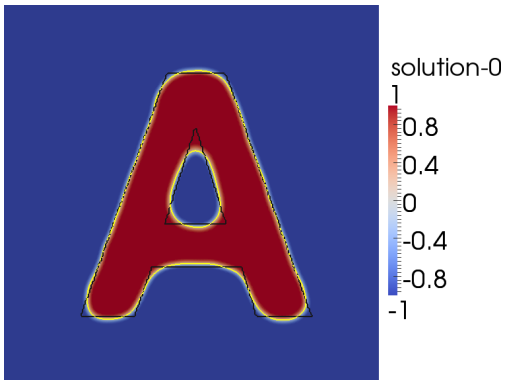
This motivates us to consider in this section how well we can recover the



(a)



(b)



(c)

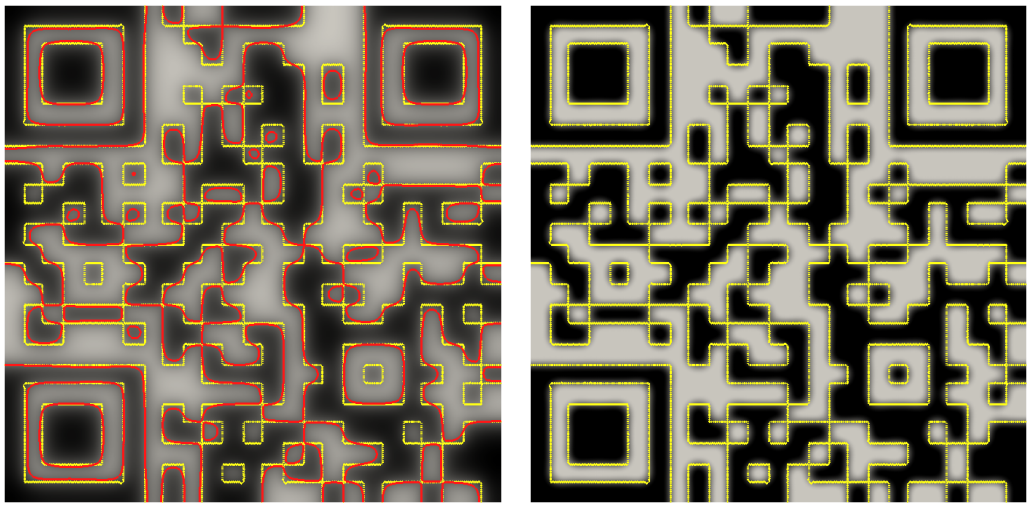
Figure 5.3: $\alpha = 0.01$, $\gamma = 0.2$, $\sigma = 1e-4$, $\varepsilon = 0.00879$ and $h = 0.00345$ using the double obstacle potential.

locations of the discontinuities with ε of moderate size (rather than as small as possible), which is computationally cheaper. In this case the choice of potential does not just affect the implementation and speed of the iterative method; the recovered functions will in general look quite different, and there may be differences in how accurately or reliably the locations of the discontinuities are recovered.

As in Section 5.6 we consider a problem with blurring caused by the solution operator of the elliptic PDE (5.5) and additive Gaussian noise of mean zero and variance γ. We use the discrete iterative method of Remark 5.20 for the smooth double well potential and (5.35) for the double obstacle potential.

At this stage it is helpful to recall the parameters we have introduced so far, as well as introduce a new parameter ω, the width of the smallest bar in the binary function. The parameters are contained in Table 5.7, and have been classified as follows:

- Problem parameters - Define the problem we are trying to solve. In appli-



(a) (b)

Figure 5.4: $\alpha = 5e-4$, $\gamma = 0$, $\sigma = 1e-5$, $\varepsilon = 0.00373$ and $h = 0.00146$ using the double obstacle potential. Here σ is smaller than in Figures 5.2 and 5.3 as the length scale of the features we are trying to resolve is smaller. Issues surrounding the choice of σ are discussed in Appendix 5.A.1.

cations we have no control over these, though we suppose they are known a priori.

- Model parameters - Specify the model we will use to solve the problem. Different values can lead to the recovery of quite different functions, so they need to be chosen carefully.
- Approximation parameters - We do not work with the model, but rather an approximation of it. These parameters control how good the approximation is.
- Discretisation parameters - Affect the accuracy of the spatial discretisation in the iterative method.
- Iteration parameters - Determine the behaviour of the iterative method.
- Implementation parameters - Control the finer details of the implementation.

We also have a number of less significant implementation parameters that handle the imprecision of computer arithmetic. These will be set to sensible values and ignored in our discussion.

Parameter	Description	Type of parameter	Optimal value
ω	Width of smallest bar in binary function	Problem	-
α	Level of blurring	Problem	-
γ	Level of noise	Problem	-
σ	Weighting given to perimeter regularisation	Model	$\omega/80$
ε	Order of width of interface	Approximation	$\omega/4\pi$
h	Grid width	Discretisation	$\omega/32$
u^0	Initial iterate	Iteration	-
ρ	Parameter in iterative method	Iteration	DW: 0.833, DO: 0.588
TOL	Stopping criterion	Implementation	DW: $3e-4$, DO: $3.5e-4$

Table 5.1: Classification of parameters.

Motivated by the above discussion we now investigate differences between the smooth double well and double obstacle potentials in accuracy, reliability, speed, and implementational complexity.

5.7.1 Accuracy

Denote the binary function we want to recover by $\bar{u}$ and the recovered function by $u_{\varepsilon,h}$. We measure the accuracy of the recovery by calculating the error quantity

$$E(u_{\varepsilon,h}) := \frac{1}{4}\big|\, |P(u_{\varepsilon,h})|_{TV} - |\bar{u}|_{TV} \,\big| + \frac{1}{2}\|P(u_{\varepsilon,h}) - \bar{u}\|_{L^1(\Omega)},$$

where P is the L^2 projection onto the space $BV(\Omega, \{-1, 1\})$ (i.e. $P(u_{\varepsilon,h}) = 1$ when $u_{\varepsilon,h} \geq 0$ and -1 when $u_{\varepsilon,h} < 0$). $|u|_{TV}$ is the total variation of u, as defined in Section 5.1.2. The integer part of $E(u_{\varepsilon,h})$ tells us the absolute difference between the number of bars in the projected recovered function and $\bar{u}$. The decimal part tells us whether the discontinuities in the projected recovered function are in the correct locations. So E measures the accuracy of the recovery in a sense that matters in applications.

We project because our best guess of $\bar{u}$ should lie in $BV(\Omega, \{-1, 1\})$. The downside of this is that $P(u_{\varepsilon,h})$ is not a minimiser of (5.7). It is important to note that the recovery using the double obstacle potential is naturally much closer to being binary than with the smooth double well potential, so projection is less

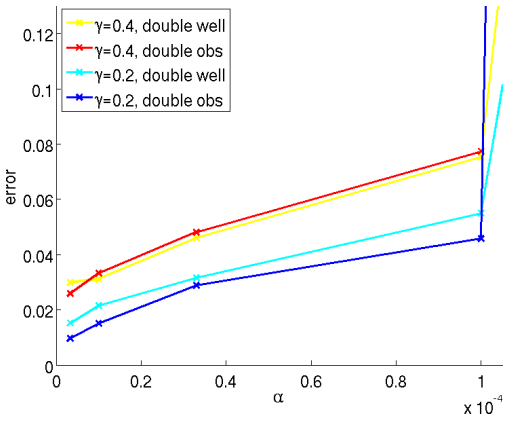


Figure 5.5: The error (averaged over different realisations of the noise) for both potentials at different levels of blurring and noise.

necessary. This is a big advantage of using the double obstacle potential, which must be remembered when values of $E(u_{\varepsilon,h})$ seem comparable.

The test problems we use for our comparison use the same binary function as in Section 5.6.1 (which has $\omega = \frac{1}{113}$), and different levels of blurring and noise i.e. a range of values of α and γ. We first fix σ based on the size of ω (as described in Appendix 5.A.1) then choose good values of the approximation and discretisation parameters (as described in Appendices 5.A.2 and 5.A.3). So we have $\sigma = 1e-4$, $\varepsilon = 7.06e-4$ and $h = 2.77e-4$ for both potentials. Each realisation of the noise will be different, so we calculate an average E over multiple realisations of the noise. As we observed earlier, we get the same steady state of (5.7) regardless of the choice of iteration parameters. The same is true for implementation parameters. So we ignore both these types of parameters in our discussion of accuracy.

We see in Figure 5.5 that neither potential is the most accurate in all circumstances. For moderate levels of noise ($\gamma = 0.2$), the double obstacle potential leads to a slightly more accurate recovery. However for high levels of noise ($\gamma = 0.4$), the smooth double well potential seems to perform slightly better. Without projection the double obstacle potential always leads to a recovery which is significantly more accurate than the smooth double well potential.

5.7.2 Reliability

By reliability we refer to the range of problems (i.e. the levels of blurring and noise) over which a binary function can be recovered with reasonable accuracy; as the amount of blurring and noise are increased, eventually the recovered function does not resemble the binary function we wanted. Note that this range will depend on

	Time for rough recovery (s)	Time for accurate recovery (s)
Smooth double well		
Average time/it	0.0359	0.181
# iterations	11	170
Runtime	0.41	29.9
Double obstacle		
Average time/it	0.0639	0.255
# iterations	9	170
Runtime	0.58	42.6

Table 5.2: Average runtimes for with $\alpha = 1e-4$ and $\gamma = 0.2$.

σ. We do not do a detailed comparison of reliability, but feel that it is comparable for both potentials. For example, we can see in Figure 5.5 that $\alpha = 1e-4$ and $\gamma = 0.4$ is roughly the limit at which the correct number of bars can be recovered using either potential.

5.7.3 Speed

The time it takes to recover a function which resembles the binary function is an important practical consideration. Where as accuracy is independent of the implementation, this is certainly not the case for speed. All but the inner workings of each iterative method in our implementation are identical, so we will do our best to make a fair comparison of speed.

We perform this comparison for the binary function of Section 5.6.1, one choice of blurring and noise ($\alpha = 1e-4$ and $\gamma = 0.2$), and σ as in Sections 5.7.1 and 5.7.2. Choices of ε and h as well as iteration and implementation parameters have a big impact on speed, so we will test two different combinations of these parameters. Our timings can be found in Table 5.2.

The runtimes for 'accurate recovery' use ε and h as in Sections 5.7.1 and 5.7.2, and TOL as described in Appendix 5.A.5. These values have been chosen to ensure robustness. The table also contains timings for 'rough recovery', where less conservative parameter values are used ($\varepsilon = \frac{\omega}{2\pi}$, $h = \frac{\omega}{20}$, and TOL as described in Appendix 5.A.5). For many problems we can still get a reasonable recovery with these parameter values, and it lowers the computation time significantly.

The recovery times are comparable for each potential for both rough and accurate recovery, though the smooth double well potential has a slight advantage for this size of problem. However we remark that the recovery time of the double obstacle potential scales better as the number of degrees of freedom in the discretisation increases, so it has better performance in 2D.

5.7.4 Implementational complexity

Implementing the iterative method for the double obstacle potential is less standard as we are solving variational inequality rather than a PDE. But it is no more complicated than implementing adaptivity, which is needed for the computational cost of the iterative method for the smooth double well potential to scale well to dimensions 2 and higher.

5.7.5 Summary of comparison

Both potentials can accurately recover binary functions over the same range of blurring and noise. If no projection is used, the double obstacle potential produces significantly more accurate results. Even with projection it is more accurate for moderate levels of blurring and noise. Our implementation using the smooth double well potential is slightly quicker for both accurate and rough binary recovery on our 1D test problem. However our implementation using the double obstacle potential, which is overall no more complicated, scales better to many degrees of freedom and so tends to be quicker in higher dimensions.

5.8 Materials science application

In this section we describe ongoing work where we apply our binary recovery algorithm to a problem in materials science. We will not introduce all the relevant science as it is beyond the scope of this thesis; instead we will view the problem as an image processing problem. This is collaborative work with the material scientist Dr Nils Warnken from the University of Birmingham, who believes our algorithm can solve a problem he faces in his research.

Dr Warnken is currently performing an experiment which causes approximately cube shaped crystals (which we refer to as structures) to form and merge in a material. He would like to understand how variables in the experiment affect the size and shape of these structures. A method for doing this is to take a slice of the material and use a scanning electron microscope (SEM) to produce a micrograph, which is effectively a blurred and noisy 'photograph' of the slice. The structures are aligned and consist mainly of cube and cuboid shaped structures. Therefore an appropriately aligned slice contains mainly square and rectangle shaped structures. A typical micrograph produced by a powerful SEM can be seen in Figure 5.6(a). Unfortunately such SEMs are expensive. A micrograph produced by the type of SEM that is often used in practice (for a different slice of material) can be seen in

Figure 5.6(b). This micrograph is of much lower quality; the density of the electron beam is low so little information is collected when scanning any given 'pixel', resulting in a low signal to noise ratio.

Dr Warnken would like to determine the structures represented by such micrographs. One end goal is to produce a histogram of the areas of the structures. Then the distribution of areas can be compared between slices from different experiments. The current methodology to achieve this is as follows:

1. Trace the structures in a number of micrographs by hand;
2. Use an algorithm in a piece of image processing software called ImageJ to calculate the areas of the traced structures;
3. Interpret the results e.g. by plotting histograms of areas for different slices.

Note that Step 1 is very time consuming and tedious, so it is not possible to perform it routinely. It is made worse by the fact that tracing is highly subjective. So for consistency it is necessary to have one person trace all the structures from hundreds of micrographs. If Step 1 could be satisfactorily automated then it would allow the interpretation of more experiments. Even if the automated process is computationally expensive and does not perform as well as manual tracing, it would likely still be preferable: In science it is advantageous to have a consistent, well defined, repeatable methodology. In this section we investigate the problem of determining the boundaries of the structures computationally, and report on our preliminary results.

We now describe some features of the micrographs in more detail:

- The micrographs have very large noise. This prevents a simple method like thresholding being effective. We show in Figure 5.7 that a simple area counting algorithm will not work on a thresholded micrograph. Therefore a more advanced method is needed. Under reasonable assumptions on the form of the structures and the blurring (which gives us a guess of $S\bar{u}$), the noise seems to resemble Gaussian noise.
- There are two types of blurring at play: One from the focus of the SEM, as part of the slice may be further away from the SEM than the rest. This blurring has a spatial variataion. The other is a local blurring effect caused by the physics of the SEM (the electrons penetrate the material and report back what is in a ball beneath it).

- In the process being investigated, the structure grow then merge with other structures. As a result the slices contain shapes other than squares, for example 'L' shapes. If some structures are part way through merging, it is subjective whether to count a feature as one structure or two separate structures.

- Recall that the structures are three dimensional. As a result there are some darker regions in the micrographs, which arise from part of the slice passing through a gap between structures. The three dimensional nature of the structures also means that if the slice is not completely aligned with the structures then a slice may contain structures with what appear to be missing corners.

5.8.1 Mathematical model

We can view a micrograph as a function $y_d \in L^2(\Omega)$ defined over a rectangular domain $\Omega \subset \mathbb{R}^2$, then model the slice that micrograph represents as a binary function. In particular a function on $BV(\Omega, \{a_0, a_1\})$ that takes the unknown values a_1 inside of the structures and a_0 in the gaps.

Recall (5.4), our model for deblurring and denoising binary images, which we developed algorithms for solving in this chapter. This model assumes that a_0, a_1 and S are known. Also recall that in our implementation and numerical experiments we took S to be a global blurring corresponding to the elliptic PDE (5.5). In comparison, motivated by our observations on the micrographs, we want to consider a situation where:

- The values a_0 and a_1 are now unknown;

- The blurring operator S is local, spatially varying and also unknown.

The parameter σ, which is related to the unknown level of the noise, is still unknown.

These differences are not insurmountable, and we will apply our algorithm to this situation without modification:

- We estimate the values of a_0 and a_1 (provisional experiments suggest these do not need to be known exactly for effective recovery);

- For the time being we will just consider a small region of a micrograph so that the blurring is approximately constant;

- As the noise is the main source of corruption in the micrographs (not the blurring), we nevertheless use the global blurring operator corresponding to the elliptic PDE (5.5), which contains a parameter α. If this model shows promise we can implement a more realistic blurring operator;

- We will experiment with a variety of parameters α in this blurring operator and also vary σ to see what allows a good recovery.

So to summarise: We will fix a_0 and a_1 and explore the two dimensional σ and α parameter space.

Remark 5.22. *Techniques for recovering a_0, a_1, α and σ at the same time as the interfaces for models related to (5.4) have been considered in the literature. We will not use these as we are currently only doing a study of the feasibility of using our algorithm for this materials science application. In particular, we just want to know whether parameter values exist that give a good recovery, and are not (yet) concerned about an automated way of finding such parameter values.*

Remark 5.23. *The term 'good recovery' in the above remark is subjective as we do not know what the slice truly looks like.*

5.8.2 2D binary recovery

We will now apply our double obstacle binary recovery algorithm in the way we just described in order to recover the structures in a subset of Figure 5.6(b). A small subset such as this decreases the computation time of the algorithms and allows for more experimentation with the parameter values σ and α. We avoided a subset with any unusual defects. We take the subset to be defined on $\Omega := (0,1)^2$ and its (scaled) interpolation onto the piecewise linear finite element space V_h with $h = 0.00521$ is shown in Figure 5.8.

Remark 5.24. *Note that the resolution of the data is quite high so perhaps some prepossessing could be performed by taking local averages. We will not do this and just use the raw data.*

For our first experiments we set $a_0 = -1$ and $a_1 = 1$. This is likely not optimal, as the scaling of the micrograph onto $[-1,1]$ can be skewed by an unusually large spike in the noise, but it is what we do initially. The smallest feature we want to resolve is the gap between the structures. So recalling that the domain of the function represented by Figure 5.8(b) is $\Omega = (0,1)^2$, we estimate that we should take $\omega = 0.5$. Now motivated by the parameter studies from Section 5.7 (which were for a 1D problem but also work well in practice for 2D problems), we take $\varepsilon = 6h$ and $h = 0.00521 = 1/192$ (i.e. $c_1 = 0.5$, $c_2 = 3$). Here we are not being conservative; we are taking the largest values we can get away with in order to get a fast computation time. Figure 5.9 shows $L^2(\Omega)$ projections onto $\{-1,1\}$ of the recovered functions for

various α and σ. We are more interested in its practical effectiveness at recovering the structures, rather than the validity of the model, which is why we do not look at unprojected minimisers of (5.4).

In Figure 5.9 we make the following observations:

- Larger σ leads to the removal of larger artifacts (small holes), which are caused by the noise. However if it is too large then more necking occurs between structures.
- Larger α has a similar effect; it encourages some of the artifacts to disappear, but can lead to necking if it is too large.
- There are no artifacts in the gaps between the object, but just in the objects themselves. This suggests that maybe level of noise is less in the channels, but for now we will not change our model to take this into account.

The artifacts in some of the recoveries are not necessarily a problem. Post-processing could be used to remove them e.g. by filling in any hole with an area less than a certain threshold. Necking is more of a problem as it could lead to two objects being considered as one large object. In some cases this is justified as two objects may be in the process of merging, as is the case in the bottom of the micrograph. Baring this in mind, Dr Warnken thought that the middle Figure 5.9(e) with $\sigma = 1e-3$ and $\alpha = 1e-4$ was the best.

We now change our choice of a_0 and a_1. If we suppose that there is no blurring ($\alpha = 0$), then by examining a slice of the data (see Figure 5.10) it seems reasonable to take $a_0 = -0.6$ and $a_1 = 0.2$. Figure 5.11 shows the result of using these values compared to our previous values. Although this removes more artifacts of the noise, it leads to more necking between objects.

These initial numerical tests show strong potential for the application of our algorithm to recovering the structures in micrographs. In particular, it removes necking effectively, so combined with pre and postproccessing it could lead to reliable automation of this task.

5.8.3 Alternative approaches

Since the structures are 3D objects, we may be able to use micrographs of multiple slices that are close together relative to the size of the structures in order to get a better recovery. This is because the noise should be independent between the different micrographs, and therefore may cancel out to some extent. We will compare:

1. The 2D recovery approach of the previous subsection (with $a_0 = -1$ and $a_1 = 1$);

2. Averaging multiple micrographs then using the 2D recovery approach;

3. Combining micrographs of multiple slices to create a 3D micrograph, then using 3D binary recovery on this. A slice can then be taken through the middle to get a 2D recovery.

We will explain these approaches in more detail shortly.

We do not currently have access to micrographs of slices that are sufficiently close together, though it is possible to collect these. Instead we artificially generate a micrograph of a key feature of the structures; a junction where three structures meet (see Figure 5.12(a)). Here the grey regions are structures and the black channel is the gap between them. We expect that if we can successfully recover such a feature, then we can likely recover all the structures well.

To artificially construct a blurred and noisy micrograph of this feature we define a function on $(0,1)^2$ which takes the value 1 on the structures in Figure 5.12(a) and -1 in the gaps. We then apply the blurring operator S (still defined as the solution operator of the elliptic PDE) on a fine triangulation for some α and add Gaussian random noise of variance γ. Such a micrograph can be seen in Figure 5.12(b) for $\alpha = 1e-2$ and $\gamma = 3$.

For our tests we construct 5 such micrographs. So we have 5 micrographs with the same underlying feature and blurring, but independent realisations of the noise. We can use these to define a 3D micrograph in the domain $\Omega = (0,1)^2 \times (0, 0.2)$; we suppose each micrograph is 0.05 units apart, and we linearly interpolate between them (see Figure 5.12(c)). Note that we are supposing that the micrographs are sufficiently close together and the underlying feature does not change between the different micrographs. This is a strong assumption and we will need to investigate how robust our approaches are when this is not the case.

We now describe the two new approaches in more detail.

Averaging approach

This involves taking a pixelwise average of multiple micrographs which gives an averaged micrograph. We then apply a 2D binary recovery to this. If the slices are close together then we hope the micrographs will represent the same features but have different realisations of the noise. So the averaging will reduce the level of the noise, enabling a better recovery. This approach is very simple and we mainly use it

as a benchmark for our 3D recovery approach. An artificially generated micrograph resulting from averaging 5 micrographs like the one in Figure 5.13(a) can be seen in Figure 5.13(b).

3D binary recovery approach

For this approach we interpolate multiple micrographs to form a 3D micrograph, scale this to the interval $[-1, 1]$, and then use this as the data for a 3D binary recovery i.e. the same approach as in Section 5.8.2, but in 3D instead of 2D. In particular, we still take $a_0 = -1$ and $a_1 = 1$. We can then take a slice of the projected recovered function to get our best guess of the boundaries of the structures. In our numerical tests we use the same 5 artificially generated micrographs as for the averaging method.

Comparison

The width of the largest feature we want to recover is 0.2, so we take $\varepsilon = 6h$ and $h = 0.0433013$ i.e. $c_1 = 0.5$ and $c_2 = 3$, as in Section 5.8.2). Our results can be seen in Figure Figure 5.14.

We find that the averaging approach improves the recovery compared to recovering an unaveraged micrograph, which is not surprising. The 3D recovery approach performs better still and the gaps are recovered accurately. This approach looks very promising provided we can take sufficiently close together slices relative to the size of the structures.

Note that if some of the slices have a different underlying structure (e.g. if instead of the bottom slice going through a structure, it instead goes through a gap), then this could damage the effectiveness of the averaging approach. The 3D binary recovery approach is likely to be robust to this. This is a direction for further research.

Further work

We finish this section with ideas for future work that could lead to better recoveries of structures. We hope to investigate these in the future:

- We should test how robust the averaging and 3D recovery approaches are to defects in one or more of the micrographs. Currently our artificial micrographs only differ in their realisations of the noise.
- We could investigate taking weighted averages of the micrographs with further away micrographs getting a lower weighting.

- It is important to test the averaging and 3D recovery approaches on real micrographs.
- We could use ideas from the anisotropic Allen-Cahn equation to perform a recovery which favours vertical and horizontal straight lines. This would take advantage of our prior knowledge of the alignment of the structures.

5.A Parameter choices

In this appendix we describe our methodology for choosing parameter values for the numerical tests and comparisons in Sections 5.6 and 5.7.

5.A.1 Choice of model parameter σ

We recover different functions for different values of σ, so it is important to choose the 'right' value. This is illustrated in Figure 5.15, where we show the recovered functions for the same problem as in Figure 5.1(a) for different values of σ. We see that $\sigma = 5e-3$ leads to too few bars being recovered. The recovered function for $\sigma = 1e-6$ follows the noise too much and does not resemble a binary function. With $\sigma = 1e-4$ we recover something close to the binary function that generated the data, so we consider this to be a good value.

It is known that the choice of σ in (5.4) should be related to the variance of the noise. Noise with a large variance requires a large σ in order for good recovery. We could try to figure out the variance of the noise from the data and use this to choose σ, however there is not an explicit form for the relationship. Instead we choose σ based on the length scale of the features that we want to recover (i.e. the parameter ω), and use the same σ for all levels of noise. In applications this is generally known a priori e.g. for barcode recovery. This approach works well because we take σ be as large as possible while not removing the features we want to recover, and hence perform the maximum amount of denoising. We do not seem to pay a significant price for this large σ in cases where the noise is small, and this approach leads to a simple rule for choosing σ. The literature that gives us a heuristic way of choosing such a σ is introduced below.

The following result shows that it is unwise to take σ too large.

Proposition 5.25. *There exists a $\sigma^* > 0$ such that the minimiser of (5.4) is 0 iff $\sigma > \sigma^*$.*

Proof. Proposition 5.7 in [Chan and Esedoglu, 2005]. □

But we also need to be careful not to take σ too small or else the problem is ill-posed. Since S is known we have the following result in the 1D case.

Theorem 5.26. *In the absence of noise there exists a $\sigma_* > 0$ such that the minimiser of (5.4) is $\bar{u}$ whenever $\sigma \leq \sigma_*$.*

Proof. Proposition 5 in [Esedoglu, 2004]. □

Another interesting result is Theorem 1.1 part 2 in [Choksi and Gennip, 2010], which proves more explicit conditions on σ to ensure exact recovery in the case that S is a convolution with a hat function in 1D. Due to our complicated form for S we are forced to use a more heuristic argument to choose a good value for σ.

[Chan et al., 2006] shows that for the 1D case in the absence of blurring and noise (i.e. binary data), local and global minimisers of (5.4) can be calculated explicitly for a given value of σ. These calculations suggest we should take σ to be smaller than a quarter of the size of the smallest object we want to recover, or else it will not appear in the minimiser. In particular, $\sigma = \frac{\omega}{8}$ seems like a reasonable choice. But this assumes binary data. We have blurring, which means the differences between the functions in the $\|Su - g_d\|^2_{L^2(\Omega)}$ term can be much smaller. Hence we take σ an order of magnitude smaller i.e. $\sigma = \frac{\omega}{80}$. This σ is still larger than the length scale of the noise (which is of order h), so the results in [Chan et al., 2006] say it will be removed. Numerical experiments confirm that this choice of σ works well in practice.

5.A.2 Choice of ε

The phase field approximation in (5.7) results in solutions with interfaces of width $o(\varepsilon)$. In order for an accurate spatial approximation we need a reasonable number of grid points across the interfaces. So a smaller ε requires more grid points and a higher computational cost. With this is mind we want to take ε as large as we can while still resolving the finest features of the binary function. So the choice of ε should be related to the value of ω.

We assume that there is a linear relationship between the optimal choice of ε and ω and deduce the constant of proportionality c_1 such that we get a good recovery with $\pi\varepsilon = c_1\omega$. Note that $\pi\varepsilon$ is the asymptotic width of the interface for minimisers of the Ginzburg-Landau functional with the double obstacle potential, and a good approximation with the smooth double well potential. The width of interfaces in minimisers of (5.7), a perturbed Ginzburg-Landau functional, are approximately the same size. So c_1 can be thought of as the relative width of the interface compared to the width of the smallest bar.

To determine c_1 we recover a simple binary function which can be seen in Figure 5.16. We take $\omega_1 = \omega_2 = \omega_3 = 0.2$ (i.e. bars of equal widths), as we found the case where all bars are at the finest length scale to be the hardest for accurate recovery. We consider different levels of blurring and noise and compute the error E of the recovered functions. We take σ to be the optimal value of $\frac{\omega}{80}$ that we decided upon in Appendix 5.A.1, and take $\pi\varepsilon = 50h$ to ensure that effects of the spatial discretisation do not distort our results.

We observe that for a high signal to noise ratio we can take c_1 very large and still get accurate recovery ($\alpha = 0.01$ in Figure 5.17), even though the bars do not separate properly (see Figure 5.18(a)). For low signal to noise ratios ($\alpha = 0.1$ in Figure 5.17) we need to take $c_1 \leq 0.5$ for accurate recovery, though it is not until $c \leq 0.25$ that the interfaces start to look reasonably sharp (see Figure 5.18(b)). As expected there is not an accuracy penalty for taking c_1 too small, however it increases computation time by forcing us to take smaller h in order to resolve the interfaces. This motivates us to take $c_1 = 0.25$ i.e. $\pi\varepsilon = \frac{\omega}{4}$.

5.A.3 Choice of h

We use the same test problems as in Appendix 5.A.2 to deduce a constant factor c_2 such that we get a good recovery with $\pi\varepsilon = c_2 h$. Hence c_2 can be thought of as the number of grid elements across each interface.

With a high signal to noise ratio ($\alpha = 0.01$ in Figure 5.19) it can actually be advantageous to have few grid points across the interface. In this case the recovered function would have to deviate a long way from the binary function in order for the projection to take an incorrect value on even a single grid point, and the data does not force sufficient deviation. As a result we can actually get perfect recovery on coarse grids. However, if we have a poorly resolved interface we are not well approximating our model and we may get a bad recovery for low signal to noise ratios ($\alpha = 0.1$ in Figure 5.19).

We do not want to adjust the relationship between ε and h for different levels of blurring and noise; we want a relationship for each potential that always works. This means we must properly resolve the interfaces. Figure 5.19 suggests that we can take $c_2 = 5$ for both potentials, however this leads to slightly jagged interfaces. Therefore we will again favour robustness and choose $c_2 = 8$ i.e. $\pi\varepsilon = 8h$.

5.A.4 Choice of iterative parameter

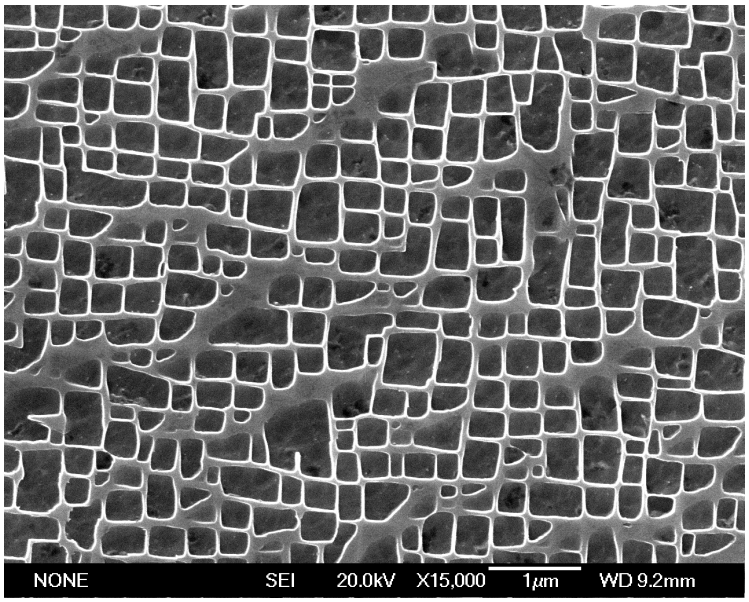
The discrete iterative methods of Section 5.5.2 have values $\bar{\rho}$ independent of h such that for all $\rho > \bar{\rho}$ the iterates decrease in energy and converge in some sense. For example, a possible $\bar{\rho}$ for the iterative method of Example 5.21 applied to the problem in Section 5.7.3 is $\max(\frac{\sigma_2}{\varepsilon}, C_s^2) = 0.999$, where we use the Poincaré constant $1/\pi$. However in practice we observe that the iterates of this method decrease in energy and converge for $\rho \geq 0.833$. It is advantageous to take ρ small, as this results in fewer iterations and uses less total computational effort. So to maximise speed we experimentally determine a value of ρ which is as small as possible while still reliably giving a decrease in energy and convergence of iterates. This approach also works for the iterative method of Remark 5.20 for the double well potential, which lies outside of our framework. So for the speed comparison in Section 5.7.3 we use $\rho = 0.833$ for the smooth double well potential and $\rho = 0.588$ for the double obstacle potential. In the rest of the simulations, where speed is less of a concern, ρ is taken large (and larger than $\bar{\rho}$ if it is known) to ensure we get the expected behaviour of the iterative methods.

5.A.5 Choice of stopping criterion

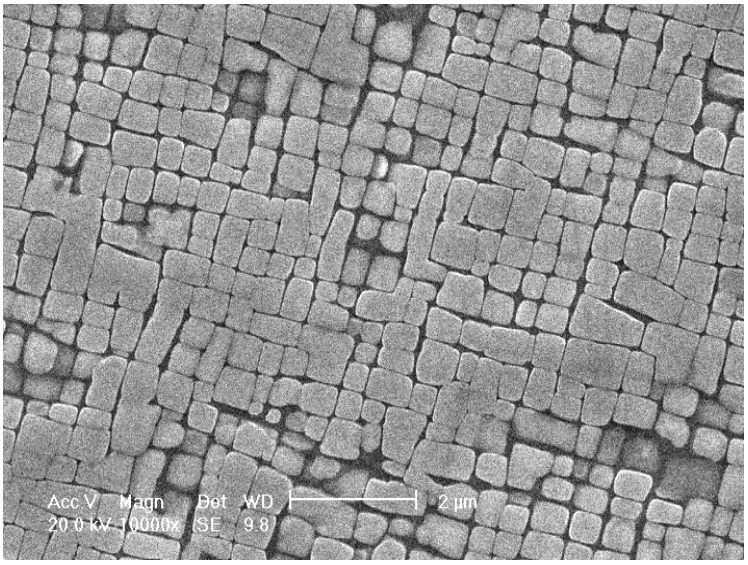
We will never quite reach the steady state of the iterative method, so a decision needs to be made about when we are sufficiently close. For this purpose we use the stopping criterion introduced in Section 5.5.3 which terminates the algorithms when the L^2 norm of the difference between consecutive iterations is less than TOL.

Mostly we take TOL small so that we are effectively finding the exact steady state, but for the comparison of speed in Section 5.7.3 we need to avoid unnecessary iterations. Figure 5.20 suggests about 170 iterations will take us quite close to the steady state for the problem under consideration. This corresponds to taking TOL=$3e-4$ for the smooth double well and TOL=$3.5e-4$ for the double obstacle, and we use these values for the 'accurate recovery'.

In practice we just want a sufficiently accurate recovery as quickly as possible. Our feeling is that the binary function is usually sufficiently accurately recovered once the error is below 0.1. At this stage the correct number of bars have formed and the locations are probably known well enough (e.g. for a different algorithm to interpret the binary function as a barcode). We see in Figure 5.20 that the smooth double well potential achieves this in around 11 iteration, which corresponds to TOL=$1.5e-2$. The double obstacle potential achieves this in around 9 iterations, which corresponds to TOL=$4e-2$. We take these values for the 'rough recovery'.



(a) Micrograph produced by a powerful SEM.



(b) Micrograph produced by a less powerful SEM on a different (but similar) slice of material.

Figure 5.6: Micrographs of similar slices of material produced by two different SEMs.

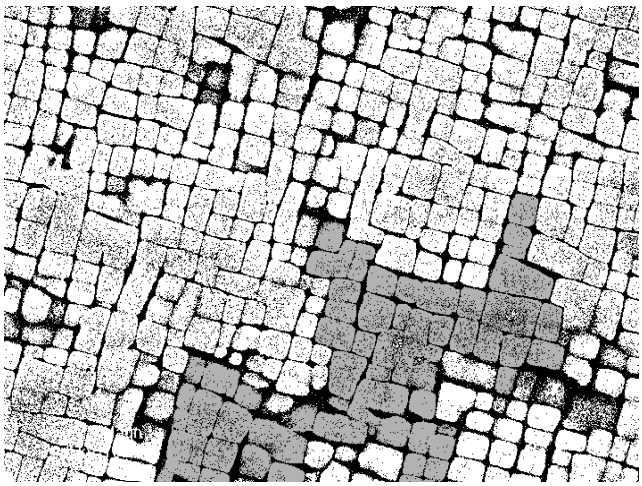
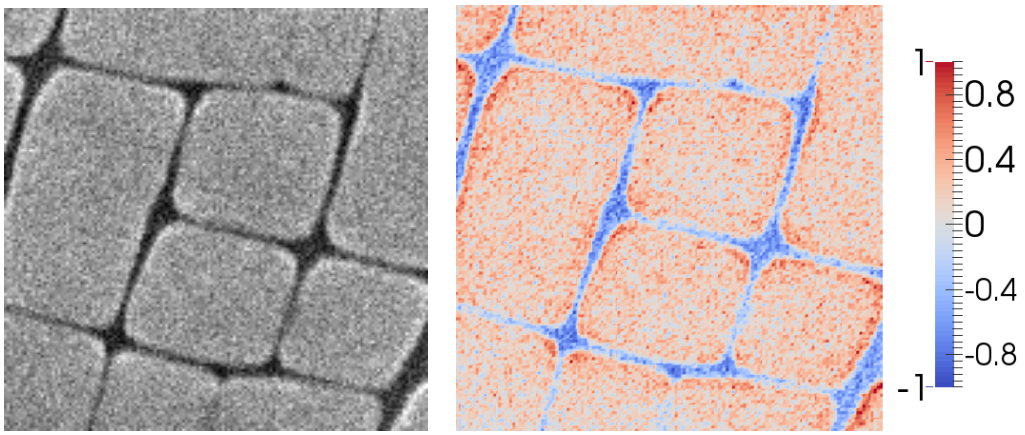


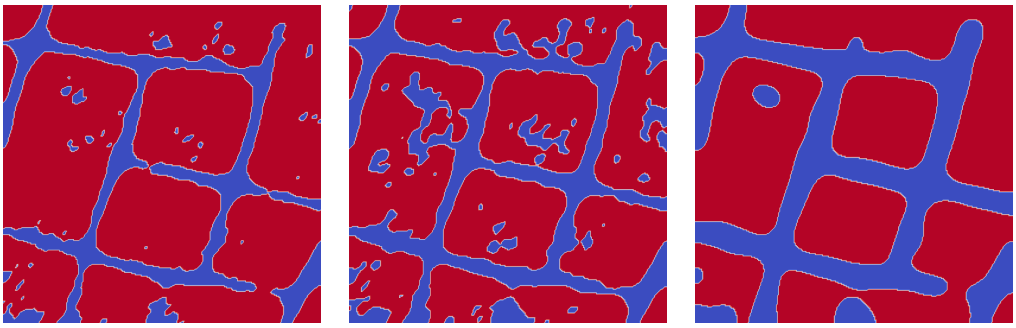
Figure 5.7: The micrograph from Figure 5.6(b) thresholded so that pixels more than half way though the greyscale spectrum are black and the others are white. The grey region highlights a connected white region, which would likely be counted as a single structure if a simple method calculating structure areas was used.



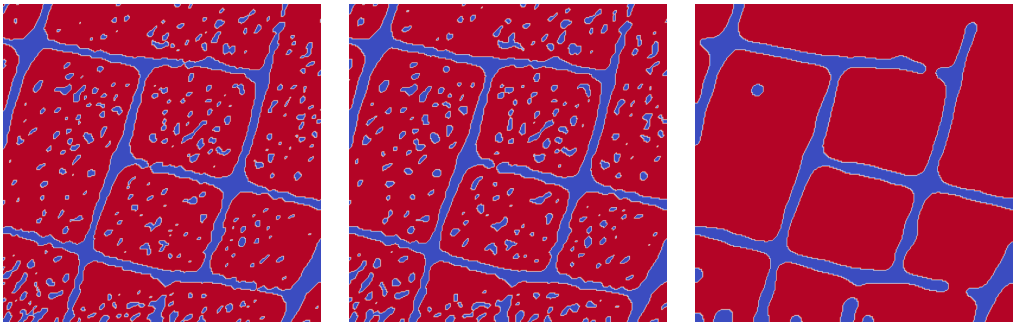
(a) Our data, which is a 150x150 pixel subimage of of Figure 5.6(b).

(b) The data from Figure 5.8(a) interpolated onto a finite element grid with $h = 0.00521$ (143648 DOFs).

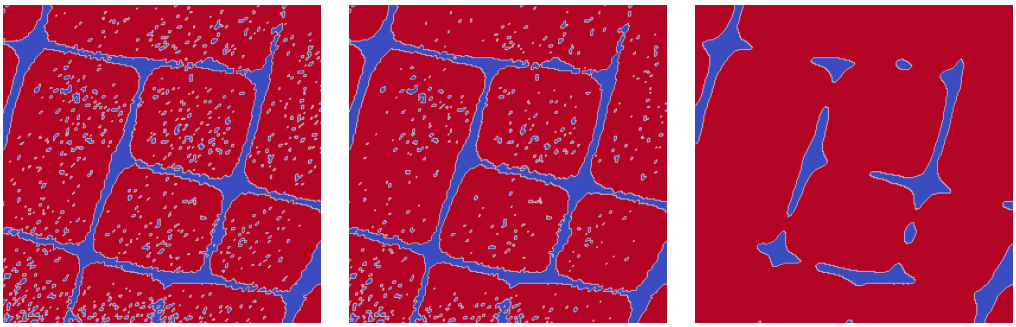
Figure 5.8: The data for binary recovery.



(a) $\sigma = 1e-4$, $\alpha = 1e-2$. (b) $\sigma = 1e-3$, $\alpha = 1e-2$. (c) $\sigma = 1e-2$, $\alpha = 1e-2$.

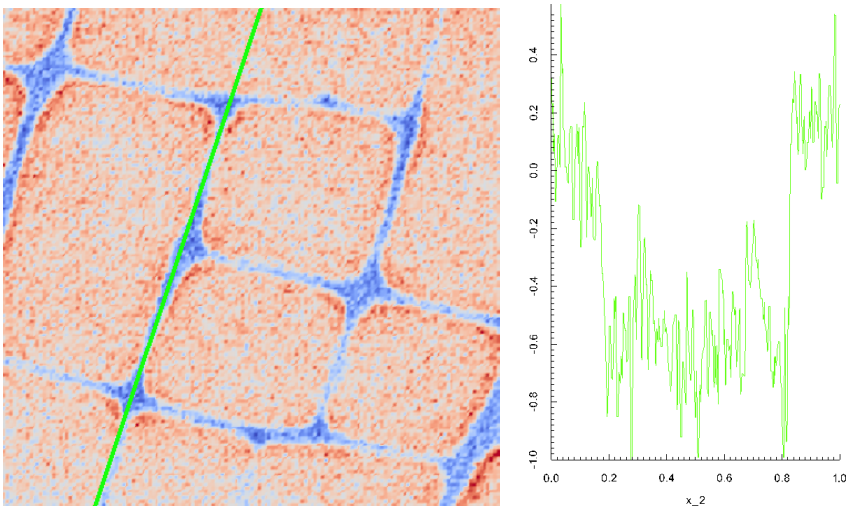


(d) $\sigma = 1e-4$, $\alpha = 1e-4$. (e) $\sigma = 1e-3$, $\alpha = 1e-4$. (f) $\sigma = 1e-2$, $\alpha = 1e-4$.



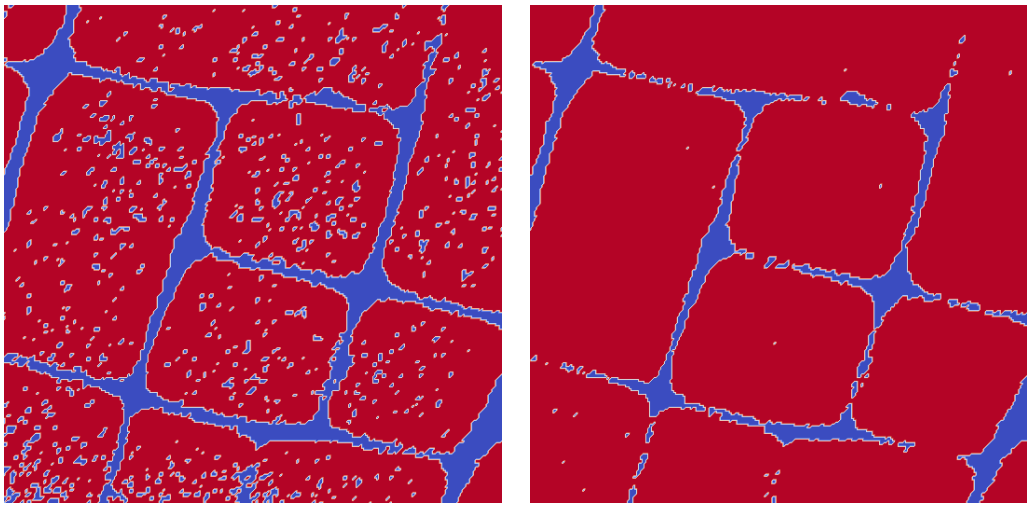
(g) $\sigma = 1e-4$, $\alpha = 0$. (h) $\sigma = 1e-3$, $\alpha = 0$. (i) $\sigma = 1e-2$, $\alpha = 0$.

Figure 5.9: Projected recoveries with different values of σ and α, with $a_0 = -1$ and $a_1 = 1$. Red regions take the value 1 and blue regions take the values -1. The boundary between these regions represent the boundaries between the structures.



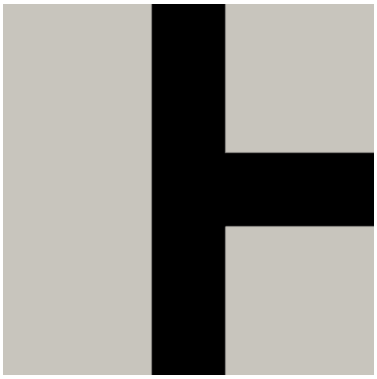
(a) (b)

Figure 5.10: Figure 5.10(b) shows the values the data takes along the green line marked in Figure 5.10(a). The suggests that the average value in the structures is around 0.2 and the average value in the gaps is -0.6.



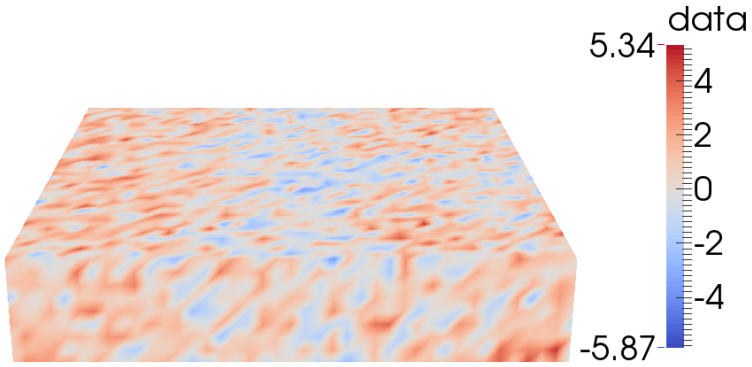
(a) $a_0 = -1$ and $a_1 = 1$. (b) $a_0 = -0.6$ and $a_1 = 0.2$.

Figure 5.11: Projected recoveries for $\sigma = 1e-4$ and $\alpha = 0$ with different a_0 and a_1.



(a) The underlying structures.

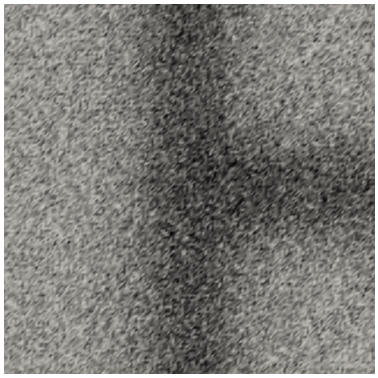
(b) One of the 5 artificially generated micrographs.



(c) The constructed 3D micrograph.

Figure 5.12: Artificially generated micrographs.

(a) Artificially generated micrograph (the same as Figure 5.12(b)).



(b) Average of 5 artificially generated micrographs.

Figure 5.13: An averaged micrograph.

(a) 2D recovery, $\alpha = 1e-3$.

(b) 2D averaged recovery, $\alpha = 1e-3$.

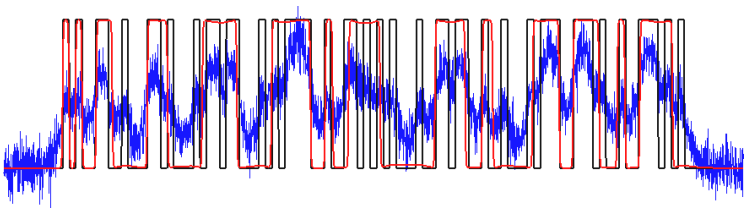
(c) 3D recovery, $\alpha = 1e-3$.

(d) 2D recovery, $\alpha = 1e-2$.

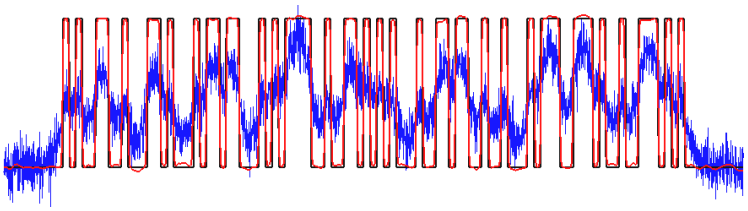
(e) 2D averaged recovery, $\alpha = 1e-2$.

(f) 3D recovery, $\alpha = 1e-2$.

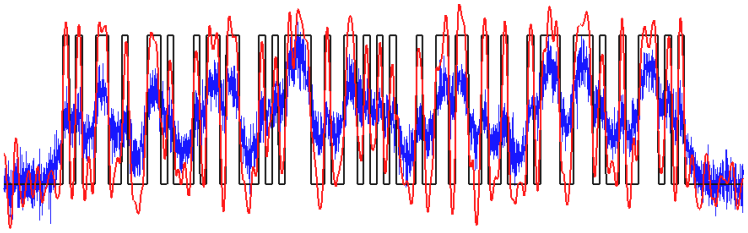
Figure 5.14: Comparison of projected recoveries using the different approaches with $\alpha = 1e-3$ (top row) and $\alpha = 1e-2$ (bottom row).



(a) $\sigma = 5e-3$.



(b) $\sigma = 1e-4$.



(c) $\sigma = 1e-6$.

Figure 5.15: The problem of Figure 5.1(a) with different values of σ.

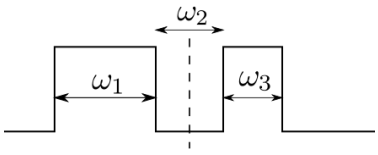


Figure 5.16: A simple binary function.

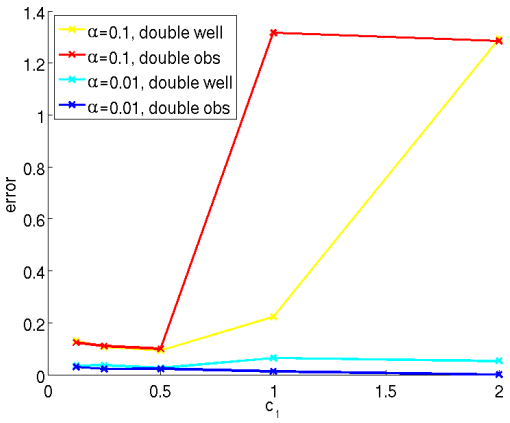
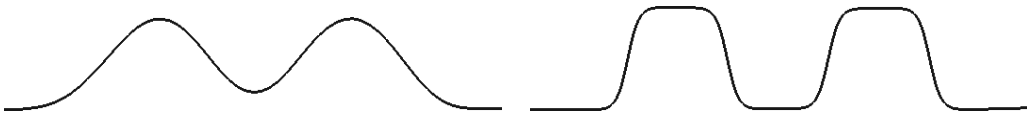


Figure 5.17: Errors (averaged over many realisations of the noise) for both potentials at different levels of blurring and $\gamma = 0.2$.



(a) $c_1 = 2.0$. (b) $c_1 = 0.25$.

Figure 5.18: The interfaces using the smooth double well potential with different values of c_1. 5.18(b) shows the interfaces for $c_1 = 0.25$, which we decide is the optimal parameter value.

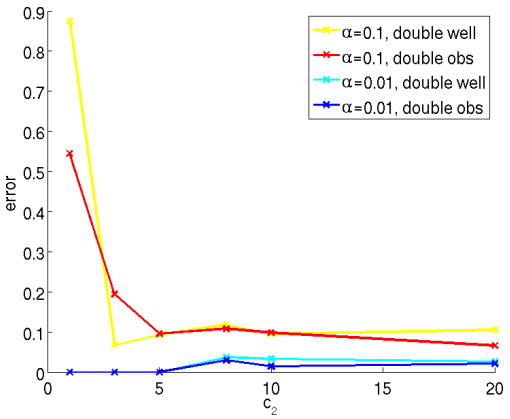
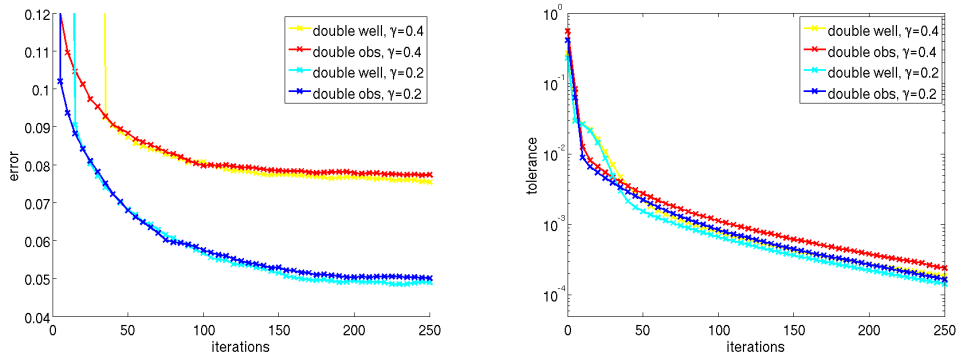


Figure 5.19: Errors (averaged over many realisations of the noise) for both potentials at different levels of blurring and $\gamma = 0.2$ with different values of c_2.



(a) Error. (b) TOL.

Figure 5.20: The error (averaged over many realisations of the noise) after a given number of iterations for both potentials for the problem of Section 5.7.3.

Bibliography

Y. Achdou. An inverse problem for a parabolic variational inequality arising in volatility calibration with American options. *SIAM Journal on Control and Optimization*, 43(5):1583–1615, 2005.

R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics*. Elsevier, second edition, 2003.

S. Agapiou, J. M. Bardsley, O. Papaspiliopoulos, and A. M. Stuart. Analysis of the Gibbs sampler for hierarchical inverse problems. *arXiv:1311.1138*, 2013.

L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*, volume 254 of *Oxford Mathematical Monographs*. Oxford University Press, 2000.

I. Babuška, J. R. Whiteman, and T. Strouboulis. *Finite elements: An introduction to the method and error estimation*. Oxford University Press, 2011.

W. Bangerth and R. Rannacher. *Adaptive finite element methods for differential equations*. Lectures in Mathematics. Birkhäuser Verlag, 2003.

V. Barbu and G. da Prato. *Optimal control of variational inequalities*. Pitman Advanced Publishing Program, 1984.

J. W. Barrett and C. M. Elliott. Finite element approximation of a free boundary problem arising in the theory of liquid drops and plasma physics. *Mathematical Modelling and Numerical Analysis*, 25(2):213–252, 1991.

P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, and O. Sander. A generic grid interface for parallel and adaptive scientific somputing. Part II: Implementation and tests in DUNE. *Computing*, 82(2–3):121–138, 2008a.

P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, M. Ohlberger, and O. Sander. A generic grid interface for parallel and adaptive scientific computing. Part I: Abstract framework. *Computing*, 82(2–3):103–119, 2008b.

P. Bastian, M. Blatt, A. Dedner, C. Engwer, J. Fahlke, C. Gräser, R. Klöfkorn, M. Nolte, M. Ohlberger, and O. Sander. DUNE Web page, 2011.

R. Becker, H. Kapp, and R. Rannacher. Adaptive finite element methods for optimal control of partial differential equations: Basic concept. *SIAM Journal on Control and Optimization*, 39(1):113–132, 2000.

O. Benedix and B. Vexler. A posteriori error estimation and adaptivity for elliptic optimal control problems with state constraints. *Computational Optimization and Applications*, 44(1):3–25, 2009.

L. Blank, H. Garcke, L. Sarbu, and V. Styles. Primal-dual active set methods for Allen-Cahn variational inequalities with nonlocal constraints. *Numerical Methods for Partial Differential Equations*, 2012.

L. Blank, M. Butz, and H. Garcke. Solving the Cahn-Hilliard variational inequality with a semi-smooth Newton method. *ESAIM: Control, Optimisation and Calculus of Variations*, 17(4):931–954, 2011.

M. Blatt and P. Bastian. The Iterative Solver Template Library. In B. Kagström, E. Elmroth, J. Dongarra, and J. Waśniewski, editors, *Applied Parallel Computing. State of the Art in Scientific Computing*, volume 4699 of *Lecture Notes in Computer Science*, pages 666–675. Springer Berlin Heidelberg, 2007.

M. Blatt and P. Bastian. On the generic parallelisation of iterative solvers for the finite element method. *International Journal of Computational Science and Engineering*, 4(1):56–69, 2008.

J. E. Blowey and C. M. Elliott. Curvature dependent phase boundary motion and parabolic double obstacle problems. In *The IMA Volumes in Mathematics and its Applications*, volume 47, pages 19–60. Springer, 1993.

J. F. Blowey and C. M. Elliott. The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy. Part II: Numerical analysis. *European J. Appl. Math.*, 3(2):147–179, 1992.

J. F. Blowey and C. M. Elliott. The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy. Part I: Mathematical analysis. *European Journal of Applied Mathematics*, 2(3):233–279, 1991.

C. Brett. *Optimal control and inverse problems involving point and line functionals and inequality constraints*. PhD thesis, University of Warwick, 2014.

C. Brett, C. M. Elliott, M. Hintermüller, and C. Löbhard. Mesh adaptivity in optimal control of elliptic variational inequalities with point-tracking of the state. *Interfaces and Free Boundaries (submitted)*, 2013.

C. Brett, A. Dedner, and C. M. Elliott. Phase field methods for binary recovery. In *Optimization with PDE constraints*, Lecture Notes in Computational Science and Engineering. Springer, 2014.

D. Calvetti and E. Somersalo. Hypermodels in the Bayesian imaging framework. *Inverse Problems*, 24(3):034013, 2008.

A. Canelas, A. Laurain, and A. A. Novotny. A new reconstruction method for the inverse potential problem. *Journal of Computational Physics*, 268:417–431, July 2014.

E. Casas, C. Clason, and K. Kunisch. Approximation of elliptic control problems in measure spaces with sparse solutions. *SIAM Journal on Control and Optimization*, 50(4):1735–1752, 2012.

E. Casas. L2 estimates for the finite element method for the Dirichlet problem with singular data. *Numerische Mathematik*, 47(4):627–632, 1985.

E. Casas. Control of an elliptic problem with pointwise state constraints. *SIAM Journal on Control and Optimization*, 24(6):1309–1318, 1986.

E. Casas and F. Tröltzsch. Error estimates for linear-quadratic elliptic control problems. *Analysis and Optimization of Differential Systems*, 121:89–100, 2003.

A. Chambolle and P. Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188, 1997.

A. Chambolle and G. D. Maso. Discrete approximation of the Mumford-Shah functional in dimension two. *ESAIM: Mathematical Modelling and Numerical Analysis*, 33(4):651–672, 1999.

T. F. Chan and S. Esedoglu. Aspects of total variation regularized L1 function approximation. *SIAM: SIAM Journal on Applied Mathematics*, 65(5):1817—-1837, 2005.

T. F. Chan, S. Esedoglu, and M. Nikolova. Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM Journal on Applied Mathematics*, 66(5):1632–1648, 2006.

X. Chen and C. M. Elliott. Asymptotics for a paraoblic double obstacle problem. *Proc. R. Soc. Lond. A*, 1994.

R. Choksi, Y. van Gennip, and A. Oberman. Anisotropic total variation regularized L1-approximation and denoising / deblurring of 2D bar codes. *Inverse Problems and Imaging*, 5(3):591 – 617, 2011.

R. Choksi and Y. V. Gennip. Deblurring of one dimensional bar codes via total variation energy minimization. *SIAM Journal on Imaging Sciences*, 3(4):735–764, 2010.

P. G. Ciarlet. *The finite element method for elliptic problems*. Studies in Mathematics and its Applications. North-Holland, 1978.

M. Crouzeix and V. Thomée. The stability in L_p and W_p^1 of the L^2-projection onto finite element function spaces. *Mathematics of Computation*, 48:521–532, 1987.

K. Deckelnick and M. Hinze. Convergence of a finite element approximation to a state-constrained elliptic control problem. *SIAM Journal on Numerical Analysis*, 45(5):1937–1953, 2007.

A. Dedner, R. Klöfkorn, M. Nolte, and M. Ohlberger. A generic interface for parallel and adaptive scientific computing: abstraction principles and the DUNE-FEM module. *Computing*, 90(3–4):165–196, 2010.

A. Dedner, R. Klöfkorn, M. Nolte, and M. Ohlberger. DUNE-FEM web page (http://dune.mathematik.uni-freiburg.de), 2011.

A. Demlow. Higher-order finite element methods and pointwise error estimates for elliptic problems on surfaces. *SIAM Journal on Numerical Analysis*, 47(2):805–827, 2009.

G. Dziuk. Finite elements for the Beltrami operator on arbitrary surfaces. In S. Hildebrandt and R. Leis, editors, *Partial Differential Equations and Calculus of Variations*, Lecture Notes in Mathematics, pages 142–155. Springer, 1988.

G. Dziuk and C. M. Elliott. Finite element methods for surface PDEs. *Acta Numerica*, 22:289–396, 2013.

C. M. Elliott and A. M. Stuart. The global dynamics of discrete semilinear parabolic equations. *SIAM Journal on Numerical Analysis*, 30(6):1622–1663, 1993.

S. Esedoglu. Blind deconvolution of bar code signals. *Inverse Problems*, 20(1): 121–135, 2004.

L. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2nd edition, 2010.

D. J. Eyre. An unconditionally stable one-step scheme for gradient systems. *Unpublished article*, 1998.

A. Gaevskaya. *Adaptive finite elements for optimally controlled elliptic variational inequalities of obstacle type*. PhD thesis, Universität Augsburg, 2013.

D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*, volume 224 of *Classics in Mathematics*. Springer, 2001.

R. Glowinski. *Numerical methods for nonlinear variational problems*. Scientific Computation. Springer, 1984.

W. Gong, G. Wang, and N. Yan. Approximations of elliptic optimal control problems with controls acting on a lower dimensional manifold. *SIAM Journal on Control and Optimization*, 52(3):2008–2035, 2014.

C. Gräser and R. Kornhuber. Multigrid methods for obstacle problems. *Journal of Computational Mathematics*, 27(1):1–44, 2009.

C. Gräser. *Convex Minimization and Phase Field Models*. PhD thesis, Freie Universität Berlin, 2011.

P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman Advanced Publishing Program, 1985.

A. Günther and M. Hinze. A posteriori error control of a state constrained elliptic control problem. *Journal of Numerical Mathematics*, 16(4):307–322, 2008.

B. Hackl. *Geometry variations, level set and phase-field methods for perimeter regularized geometric inverse problems*. PhD thesis, Johannes Kepler Universität Linz, 2006.

J. K. Hale. *Asymptotic behavior of dissipative systems*, volume 25 of *Mathematical Surveys and Monographs*. American Mathematical Society, 1988.

M. Hintermüller and R. H. W. Hoppe. An a posteriori error analysis of adaptive finite element methods for distributed elliptic control problems with control constraints. *ESAIM: Control, Optimisation and Calculus of Variations*, 14(3):540–560, 2008a.

M. Hintermüller and R. H. W. Hoppe. Goal-oriented adaptivity in pointwise state constrained optimal control of partial differential equations. *SIAM Journal on Control and Optimization*, 48(8):5468–5487, 2010a.

M. Hintermüller and R. H. W. Hoppe. Goal oriented mesh adaptivity for mixed control-state constrained elliptic optimal control problems. *Applied and Numerical Partial Differential Equations*, 2010b.

M. Hintermüller and R. H. W. Hoppe. Goal-oriented adaptivity in control constrained optimal control of partial differential equations. *SIAM Journal on Control and Optimization*, 47(4):1721–1743, 2008b.

M. Hintermüller and I. Kopacka. Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm. *SIAM Journal on Optimization*, 20(2):868–902, 2009.

M. Hintermüller and I. Kopacka. A smooth penalty approach and a nonlinear multigrid algorithm for elliptic MPECs. *Computational Optimization and Applications*, 50(1):111–145, 2011.

M. Hintermüller and A. Laurain. Electrical impedance tomography: From topology to shape. *Control and Cybernetics*, 37(4):913–933, 2008.

M. Hintermüller, R. H. W. Hoppe, and C. Löbhard. A dual-weighted residual approach to goal-oriented adaptivity for optimal control of elliptic variational inequalities. *(preprint)*, 2013.

M. Hinze. A variational discretization concept in control constrained optimization: The linear-quadratic case. *Computational Optimization and Applications*, 30(1): 45–61, 2005.

M. Hinze, R. Pinnau, and M. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, 2009.

I. Hlaváček, I. Bock, and J. Lovíšek. Optimal control of a variational inequality with applications to structural analysis. II. Local optimization of the stress in a beam. III. Optimal design of an elastic plate. *Applied Mathematics and Optimization*, 13 (1):117–136, 1985.

R. H. W. Hoppe and M. Kieweg. A posteriori error estimation of finite element approximations of pointwise state constrained distributed control problems. *Journal of Numerical Mathematics*, 17(3):219–244, 2009.

R. H. W. Hoppe and M. Kieweg. Adaptive finite element methods for mixed control-state constrained optimal control problems for elliptic boundary value problems. *Computational Optimization and Applications*, 46(3):511–533, 2010.

R. H. W. Hoppe, Y. Iliash, C. Iyyunni, and N. H. Sweilam. A posteriori error estimates for adaptive finite element discretizations of boundary control problems. *Journal of Numerical Mathematics*, 14(1):57–82, 2006.

J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*, volume 160 of *Applied Mathematical Sciences*. Springer, 2005.

K. Kunisch and X. Pan. Estimation of interfaces from boundary measurements. *SIAM Journal on Control and Optimization*, 32(6):1643–1674, 1994.

D. Leykekhman, D. Meidner, and B. Vexler. Optimal error estimates for finite element discretization of elliptic optimal control problems with finitely many pointwise state constraints. *Computational Optimization and Applications*, 55(3):769–802, 2013.

R. Li, W. Liu, H. Ma, and T. Tang. Adaptive finite element approximation for distributed elliptic optimal control problems. *SIAM Journal on Control and Optimization*, 41(5):1321–1349, 2002.

W. Liu and N. Yan. A posteriori error estimates for distributed convex optimal control problems. *Advances in Computational Mathematics*, 15(1-4):285–309, 2001.

Z. Luo, J. Pang, and D. Ralph. *Mathematical programs with equilibrium constraints*. Cambridge University Press, 1996.

F. Mignot. Contrôle dans les inéquations variationelles elliptiques. *Journal of Functional Analysis*, 22(2):130–185, 1976.

L. Modica and S. Mortola. Un esempio di Γ-convergenza. *Bollettino dell'Unione Matematica Italiana*, 14-B(5):285–299, 1977.

V. A. Morozov. On the solution of functional equations by the method of regularization. *Soviet Mathematics - Doklady*, 7(1):414–417, 1966.

C. B. Morrey Jr. *Multiple integrals in the calculus of variations*, volume 130 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1966.

D. B. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42(5):577–685, 1989.

P. Neittaanmäki, J. Sprekels, and D. Tiba. *Optimization of elliptic systems: Theory and applications*. Springer Monographs in Mathematics. Springer, 2006.

S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489, 2005.

J. Outrata, M. Kocvara, and J. Zowe. *Nonsmooth approach to optimization problems with equilibrium constraints: theory, applications and numerical results*, volume 28 of *Nonconvex Optimization and Its Applications*. Springer, 1998.

J. Outrata, J. Jarušek, and J. Stará. On optimality conditions in control of elliptic variational inequalities. *Set-Valued and Variational Analysis*, 19(1):23–42, 2011.

N. Petra and G. Stadler. Model variational inverse problems governed by partial differential equations. Technical Report ADA555315, University of Texas at Austin, Institute for Computational Engineering and Sciences, 2011.

K. Pieper and B. Vexler. A priori error analysis for discretization of sparse elliptic optimal control problems in measure space. *SIAM Journal on Control and Optimization*, 51(4):2788–2808, 2013.

R. Rannacher and R. Scott. Some optimal error estimates for piecewise linear finite element approximations. *Mathematics of Computation*, 38(158):437–445, 1982.

S. I. Repin. *A posteriori estimates for partial differential equations*, volume 4 of *Radon Series on Computational and Applied Mathematics*. Walter de Gruyter, 2008.

J.-F. Rodrigues. *Obstacle problems in mathematical physics*, volume 134 of *Mathematics Studies*. North Holland, 1987.

A. Rösch and D. Wachsmuth. A-posteriori error estimates for optimal control problems with state and control constraints. *Numerische Mathematik*, 120(4):733–762, 2012.

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

W. Rudin. *Real and complex analysis*. Tata McGraw-Hill Education, internatio edition, 1987.

L. Sarbu. *Primal-dual active set methods for Allen-Cahn variational inequalities*. PhD thesis, University of Sussex, 2010.

H. Scheel and S. Scholtes. Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity. *Mathematics of Operations Research*, 25(1):1–22, 2000.

R. Scott. Finite element convergence for singular data. *Numerische Mathematik*, 21(4):317–327, 1973.

R. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Mathematics of Computation*, 54(190):483–493, 1990.

J. R. Shewchuk. Triangle web page (http://www.cs.cmu.edu/∼quake/triangle.html).

X.-C. Tai and T. F. Chan. A survey on multiple level set methods with applications for identifying piecewise constant functions. *International Journal of Numerical Analysis and Modeling*, 1(1):25–47, 2004.

X.-C. Tai and H. Li. A piecewise constant level set method for elliptic inverse problems. *Applied Numerical Mathematics*, 57(5-7):686–696, May 2007.

F. Tröltzsch. *Optimal control of partial differential equations: Theory, methods and applications*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, 2010.

M. Ulbrich. Semismooth Newton methods for operator equations in function spaces. *SIAM Journal on Optimization*, 13(3):805–841, 2002.

A. Unger and F. Tröltzsch. Fast solution of optimal control problems in the selective cooling of steel. *ZAMM Journal of Applied Mathematics and Mechanics*, 81(7): 447–456, 2001.

R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. Wiley-Teubner, 1996.

B. Vexler and W. Wollner. Adaptive finite elements for elliptic optimization problems with control constraints. *SIAM Journal on Control and Optimization*, 47(1): 509–534, 2008.

Y.-W. Wen and R. H. Chan. Parameter selection for total-variation-based image restoration using discrepancy principle. *Image Processing, IEEE Transactions on*, 21(4):1770–1781, 2012.

W. P. Ziemer. *Weakly differentiable functions: Sobolev spaces and functions of bounded variation*, volume 120 of *Graduate Texts in Mathematics*. 1989.